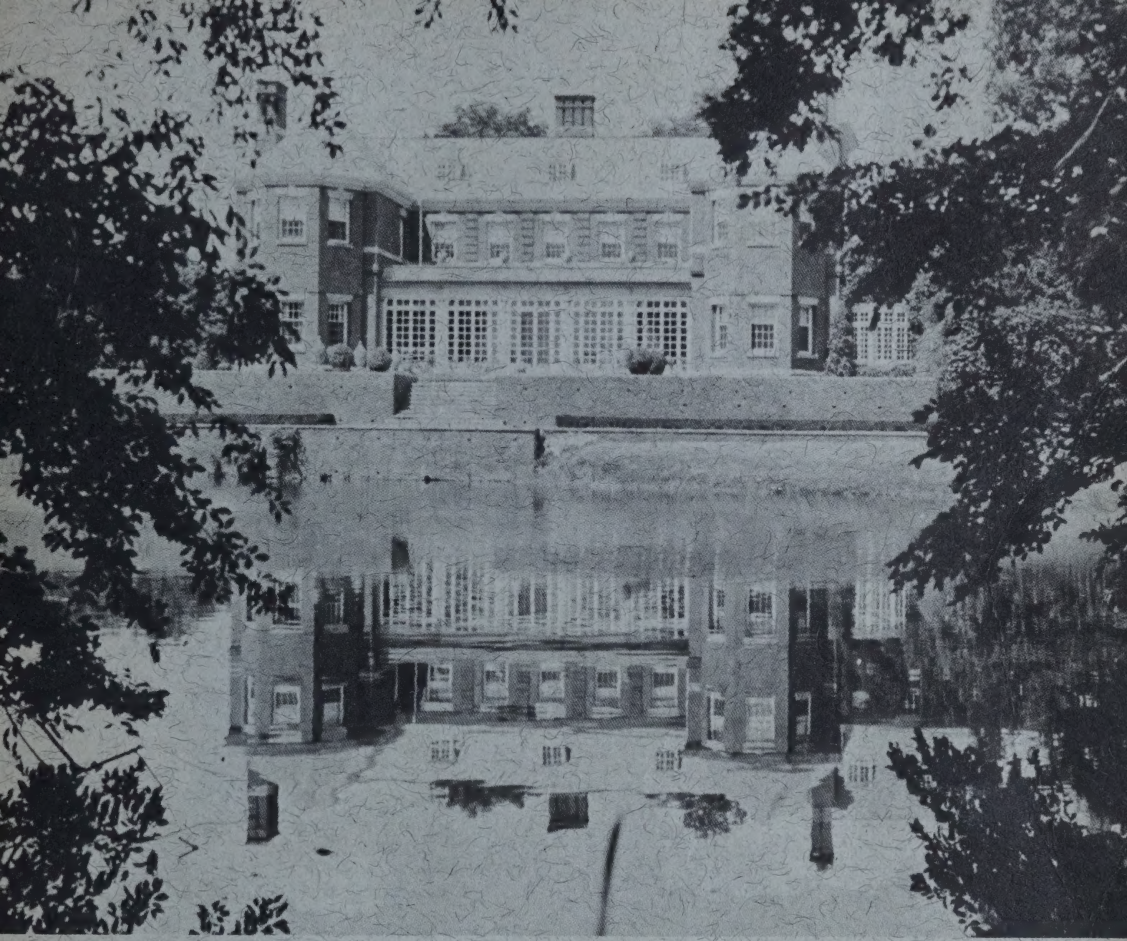


Proceedings

ELEVENTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

October 3-5, 1973

[redacted]

Allerton House, Monticello, Illinois
Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois at Urbana-Champaign

[redacted]

Allerton Conference

11th, on CIRCUIT AND SYSTEM THEORY
1973

copy 2

DEMCO

108

D. H. LAWRIE

[redacted]
LIBRARY

PROCEEDINGS

ELEVENTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

W. R. Perkins
P. V. Kokotović

Conference Co-Chairmen

Conference held

October 3, 4, 5, 1973
Allerton House
Monticello, Illinois

Sponsored by the

The Department of Electrical Engineering
and
The Coordinated Science Laboratory
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign

FOREWORD

The Eleventh Annual Allerton Conference on Circuit and System Theory was held at Allerton House on October 3-5, 1973. The Conference was co-sponsored by The Department of Electrical Engineering and The Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign. These PROCEEDINGS contain written versions of the papers presented at the Conference.

In addition to the large number of submitted papers, there were several organized sessions of invited papers. These sessions were: Infinite Dimensional Linear Systems, organized by R. W. Brockett; Power Systems, organized by A. H. El-Abiad; Socio-Economic Systems, organized by M. H. Mickle and W. G. Vogt; Games and Decisions, organized by M. Mintz and M. Simaan; Computer Architecture, organized by E. S. Davison; and Singular Perturbations, organized by P. Sannuti. We are grateful to the organizers for their time and effort on behalf of the Conference. We express our thanks to the session chairmen and to all the authors for their contributions to the Allerton Conference.

Dean E. S. Kuh presented an informative plenary address entitled "Education in the People's Republic of China". J. B. Thomas presented a stimulating talk to the Chicago, Central Illinois, and Indiana joint chapter meeting of the IEEE Group on Information Theory. This meeting was arranged by S. Tsai. We are indebted to these people for their efforts.

The support and encouragement of Professor R. T. Chien, Director of the Coordinated Science Laboratory, and Professor E. C. Jordan, Head of the Department of Electrical Engineering, are gratefully acknowledged. We thank our colleagues for their generous assistance: C. L. Chen, R. T. Chien, J. B. Cruz, Jr., E. S. Davidson, A. Haddad, W. Mayeda, G. Metze, M. Mintz, M. Simaan, T. N. Trick, and N. Wax.

We are grateful to Harold Lawler for his handling of the business aspects of the Conference. We thank Rose Harris and Rosa Townsend for their help. We appreciate the work of D. Barbour and C. Youngblut in coordinating the projection equipment. Special thanks are due Mrs. Dee Keel for her expert and skillful coordination of the Conference arrangements, and for many other invaluable contributions to the Allerton experience.

W. R. Perkins
P. V. Kokotović

TABLE OF CONTENTS

Foreword iii

I-A: INFINITE DIMENSIONAL LINEAR SYSTEMS I
Chairman and Organizer: Roger W. Brockett, Harvard University

ON CANONICAL REALIZATIONS WITH UNBOUNDED INFINITESIMAL GENERATORS 1
J. S. Baras, University of Maryland

THE REGULATOR PROBLEM AND STABILIZATION OF INFINITE DIMENSIONAL CONTROL SYSTEMS 11
R. Datko, Georgetown University

MODELS AND CRITERIA IN DISTRIBUTED STABILITY PROBLEMS . . 20
J. H. Davis, Queen's University, Canada

ON A CLASS OF MULTI-INPUT/MULTI-OUTPUT CONTROL SYSTEMS IN A HILBERT SPACE 30
P. A. Fuhrmann, Harvard University

CANONICAL FORMS IN LINEAR SYSTEMS 36
B. Moore, III, University of New Hampshire

I-B: POWER SYSTEMS
Chairman and Organizer: A. H. El-Abiad, Purdue University

TRENDS AND PHILOSOPHIES OF POWER SYSTEMS CONTROL 45
O. I. Elgerd, University of Florida

TRENDS IN STABILITY ANALYSIS OF POWER SYSTEMS 54
A. A. Fouad, Iowa State University

DATA BASE FOR REAL-TIME POWER SYSTEM CONTROL 64
A. S. Debs, Georgia Institute of Technology

I-C: ECONOMIC SYSTEMS
Chairman: J. B. Cruz, Jr., University of Illinois

A THEOREM ON CONSEQUENCES OF PLANNED DEMAND CHANGES IN A PARTIALLY FIXED-PRICE ENVIRONMENT 77
I. W. Sandberg, Bell Laboratories, Murray Hill, New Jersey

PRODUCTION CONFIGURATIONS, TECHNOLOGY SETS, AND DISCRETE SYSTEMS 82
J. A. Ortega, COPPE, Universidade Federal, Brasil and R. J. Leake, University of Notre Dame

A DYNAMIC SALES MODEL FOR REUSABLE COMMODITIES 83
J. G. Schatz, Bell Laboratories, Whippany, New Jersey

I-D: CIRCUITS
Chairman: J. J. Bourquin, University of Missouri, Rolla

STABILITY CONSIDERATIONS IN SWITCHING VOLTAGE REGULATORS 86
J. R. Wood, Texas Instruments Incorporated, Dallas, Texas

ON THE HYSTERESIS OF THE TTL CIRCUIT WITH TOTEM-POLE OUTPUT 96
R. Srivivasan and T. J. Chaney, Washington University

AN AUTOMATIC EQUALIZER FOR TV CHANNELS 106
S. Shida and N. Furuya, Nippon Electric Company, Ltd., Japan

II-A: INFINITE DIMENSIONAL LINEAR SYSTEMS II
Chairman and Organizer: R. W. Brockett, Harvard University

AN INFINITE DIMENSIONAL LOW-NOISE DIFFERENTIATOR 115
R. D. Brown and R. W. Newcomb, University of Maryland

THE 'OBSERVATION AND PREDICTION' PROBLEM FOR DIFFUSION PROCESSES 123
T. I. Seidman, University of Maryland

OPERATOR THEORY TECHNIQUES FOR DISTRIBUTED SYSTEMS 131
J. W. Helton, State University of New York

EXISTENCE OF OPTIMAL CONTROLS FOR NONLINEAR DISTRIBUTED PARAMETER SYSTEMS 141
M. Slemrod, Rensselaer Polytechnic Institute

A REALIZATION CONDITION FOR INFINITE DIMENSIONAL BILINEAR SYSTEMS 151
G. Koch, University of Rome, Italy

II-B: SOCIO-ECONOMIC SYSTEMS
Organizers and Co-Chairmen: M. H. Mickle and W. G. Vogt, University of Pittsburgh

POPULATION PROJECTIONS AND ECONOMIC INTERACTIONS 160
W. D. O'Neill, University of Illinois at Chicago Circle

INTERACTIONS BETWEEN POPULATION, PRODUCTION, AND MATERIAL BENEFIT SECTORS IN AN ECONOMETRIC MODEL OF JAPAN 171
K. Uno, Japan Economic Research Center, Japan and A. I. Duchan, Canisius College, New York

AN APPLICATION OF THE LINEAR-QUADRATIC REGULATOR PROBLEM TO AN ECONOMIC SYSTEM 178
K. Amir-Ebrahimy and W. G. Vogt, University of Pittsburgh

CALIBRATION OF AN URBAN SIMULATION MODEL ACROSS CITIES. . 187
R. B. Parker, University of Illinois, Urbana

SOCIO-ECONOMIC INFLUENCES AND INTERACTIONS WITH TRANSPORTATION SYSTEMS 194
H. Aldermeshian and M. H. Mickle, University of Pittsburgh

II-C: ESTIMATION AND CONTROL
Chairman: W. M. Wonham, University of Toronto, Canada

OPTIMIZATION OF STOCHASTIC LINEAR SYSTEMS WITH ADDITIVE MEASUREMENT AND PROCESS NOISE USING EXPONENTIAL PERFORMANCE CRITERIA 204
J. Speyer, The Charles Stark Draper Laboratory, Inc., Cambridge, Massachusetts and J. Deyst, Massachusetts Institute of Technology

ON MINIMAX PARAMETER ESTIMATION UNDER EXPONENTIATED GENERALIZED QUADRATIC LOSS 215
D. Stelte and M. Mintz, University of Illinois, Urbana

ESTIMATION OF INITIAL CONDITIONS IN LINEAR SYSTEMS . . . 226
J. S. Meditch, University of California at Irvine and G. H. Hostetter, California State University, Long Beach

ON THE SOLUTION OF A MINIMAX TERMINAL STATE ESTIMATION PROBLEM--A HILBERT SPACE APPROACH 236
G. Hexner and M. Mintz, University of Illinois, Urbana

ON-LINE REGULARIZATION IN INVERSE PROBLEMS 247
R. Rutman, Tel-Aviv University, Israel

DECENTRALIZED STOCHASTIC CONTROL OF AN INTERCONNECTED ELECTRIC ENERGY SYSTEM 254
E. C. Tacker, Louisiana State University, C. W. Sanders, The University of Wisconsin at Madison, D. R. Fischer, Louisiana State University, and T. D. Linton, The University of Wisconsin at Madison

II-D: NETWORKS I
Chairman: J. Vlach, Waterloo University, Canada

THE INTERRELATED TREE ANATOMIES OF LADDERS, WHEELS, DRUMS AND TOPS 257
B. R. Myers, University of Notre Dame

HYPERGRAPH DECOMPOSITION WITH MINIMAL CONNECTIONS . . . 266
E. Lodi and F. Luccio, University of Pisa, Italy

A COMPARATIVE STUDY OF TOPOLOGICAL ANALYSIS METHODS FOR CIRCUITS AND SYSTEMS 275
A. Y. Lee, Michigan Technological University and H. D'Angelo, Memphis State University

PASSIVE NONLINEAR NETWORKS 284
R. Saeks and L. Jenkins, University of Notre Dame and L. Winslow, Wright State University

SWITCHED ELECTRICAL NETWORKS AND BILINEAR EQUATIONS . . 294
J. R. Wood, Texas Instruments Incorporated, Dallas, Texas

III-A: GAMES AND DECISIONS I
Organizers: M. Mintz and M. Simaan, University of Illinois, Urbana
Chairman: M. Mintz, University of Illinois, Urbana

A GENERALIZED REPRESENTATION THEOREM WITH APPLICATIONS TO ESTIMATION AND HYPOTHESIS TESTING 304
K-P. Dunn and I. B. Rhodes, Washington University

STALEMATE AND OTHER PROBLEMS OF CONTROLLERS IN CONFLICT 315
D. L. Lukes, University of Virginia

MINIMAX ADMISSIBLE ESTIMATORS ON THE MEAN OF A MULTIVARIATE NORMAL DISTRIBUTION WITH NON-IDENTITY COVARIANCE MATRIX 324
M. E. Bock, University of Illinois, Urbana

THE CHARACTERISTIC FUNCTION AND THE CORE OF A MULTI-STAGE GAME OF COALITIONS 333
A. Haurie, Ecole Des Hautes Etudes Commerciales, Montreal, Canada

III-B: CONTROL SYSTEMS I
Chairman: J. L. Melsa, University of Notre Dame

THE FEEDFORWARD AND FEEDBACK CONTROL OF A GENERAL SERVOMECHANISM PROBLEM, Part I 343
E. J. Davison, University of Toronto, Canada

THE FEEDFORWARD AND FEEDBACK CONTROL OF A GENERAL SERVOMECHANISM PROBLEM, Part II 353
E. J. Davison, University of Toronto, Canada

DESIGN OF LOW-ORDER, DYNAMIC, FEEDBACK COMPENSATORS VIA A POLE-PLACEMENT ALGORITHM 363
J. R. Anders and S. H. Wu, Marquette University

A SEQUENTIAL ALGORITHM FOR CALCULATION OF OPTIMAL OUTPUT-CONSTRAINED FEEDBACK REGULATORS 373
S. P. Bingulac, Institute "B. Kidrič", Belgrade, Yugoslavia and N. M. Ćuk and M. S. Ćalović, Serbian Electric Power Company, Belgrade, Yugoslavia

THE OPTIMAL LINEAR REGULATOR WITH APERIODIC SAMPLING TIMES 383
R. A. Schlueter, Michigan State University

III-C: DIGITAL SYSTEMS I
Chairman: G. Metze, University of Illinois, Urbana

THE SOLUTION OF BOOLEAN EQUATIONS IN DESIGNING SWITCHING NETWORKS 393
J. W. Gault, North Carolina State University

A COST ANALYSIS OF HAZARD-FREE NETWORKS SYNTHESIZED FROM MONOTONE JUMP FUNCTIONS 400
E. G. DuCasse, Brooklyn College and G. A. Metze, University of Illinois, Urbana

ON THE DIAGNOSIS OF MULTIPLE OUTPUT CIRCUITS 408
D. R. Schertz, Bradley University

ON THE DETECTION OF BRIDGE AND SHORTED DIODE FAULTS . . . 412
S. M. Reddy, University of Iowa

AN APPROACH FOR SYNTHESIZING TRANSITION LOGIC CIRCUITS 418
M. J. Stucki, Washington University

III-D: ESTIMATION AND IDENTIFICATION
Chairman: L. Shaw, Polytechnic Institute of New York

LINEAR SYSTEM IDENTIFICATION: THE APPLICATION OF LION'S IDENTIFICATION SCHEME TO A THIRD ORDER SYSTEM WITH NOISY INPUT-OUTPUT MEASUREMENTS 428
C. M. Brown, Jr., The Analytic Sciences Corporation, Reading, Massachusetts and R. V. Monopoli, University of Massachusetts

ON MINIMAL ERROR ENTROPY SYSTEM IDENTIFICATION 437
P. Kalata, Illinois Institute of Technology and R. Priemer, University of Illinois at Chicago Circle

OBSERVER-ESTIMATORS FOR DISCRETE RANDOM PARAMETER SYSTEMS 447
C. S. Sims, Oklahoma State University

DISTRIBUTED PARAMETER IDENTIFICATION USING GOURSAT THEORY 453
R. O. Hughes, Construction Engineering Research Laboratory, Champaign, Illinois

APPLICATION OF STOCHASTIC APPROXIMATION TO PROCESS CONTROL 454
T. Prasad, University of Waterloo, Canada and N. K. Sinha, McMaster University, Canada

UNBIASED ESTIMATES FOR ON-LINE SYSTEM IDENTIFICATION . . 456
A. Sen and N. K. Sinha, McMaster University, Canada

III-E: SYSTEM THEORY
Chairman: J. Meditch, University of California, Irvine

INVERTIBILITY CONDITIONS FOR A CLASS OF FINITE STATE SYSTEMS EVOLVING ON GROUPS 459
A. S. Willsky, Massachusetts Institute of Technology

SYNTHESIS OF MULTIVARIABLE REGULATORS 468
B. Francis, O. A. Sebakhy and W. M. Wonham, University of Toronto, Canada

ON STEADY-STATE CHARACTERISTICS OF LINEAR MULTIVARIABLE SYSTEMS 477
S. Hosoe and M. Ito, Nagoya University, Japan

A NEW APPROACH TO SYSTEM REALIZATION FROM INPUT OUTPUT DATA 487
R. D. Gupta and F. W. Fairman, Queen's University, Canada

A UNIFIED THEORY OF THE HAMILTONIAN FORMULATION OF SYSTEMS 495
Hsien-Lu Huang, West Virginia University

IV-A: GAMES AND DECISIONS II

Organizers: M. Mintz and M. Simaan, University of Illinois, Urbana
Chairman: M. Simaan, University of Illinois, Urbana

THE ANALYSIS OF THE BORDER VIOLATION PROBLEM AS A POINT PROCESS GAME 505
A. Ephremides, University of Maryland

SMALL SAMPLE ESTIMATION FROM INFORMATION THEORETIC METHODS 515
R. L. Kashyap, Purdue University

THE GENERATION AND APPLICATION OF CAPTURE REGIONS IN AIRCRAFT/MISSILE ENGAGEMENTS 526
David M. Salmon and L. Meier, III, Systems Control, Inc., Palo Alto, California

DIFFERENTIAL GAMES WITHOUT PURE STRATEGY SADDLE POINT SOLUTIONS 536
W. E. Schmitendorf, Northwestern University

CONDITIONS FOR PARETO OPTIMALITY IN A MULTICRITERION SYSTEM WITH NONDIFFERENTIABLE FUNCTIONS 544
J.-L. Goffin, Ecole des Hautes Etudes Commerciales, Montreal, Canada

IV-B: COMPUTER ARCHITECTURE

Chairman and Organizer: E. S. Davidson University of Illinois, Urbana

COST-EFFECTIVE DESIGN OF SPECIAL-PURPOSE PROCESSORS: A FAST FOURIER TRANSFORM CASE STUDY 547
A. G. Larson, Stanford Research Institute, Menlo Park, California and E. S. Davidson, University of Illinois, Urbana

CONTROL OF A VARIABLE CONFIGURATION PIPELINED ARITHMETIC UNIT 558
C. M. Stephenson, Texas Instruments Incorporated, Austin, Texas

MEMORY SYSTEMS FOR PARALLEL ARRAY PROCESSORS 568
D. H. Lawrie, University of Illinois, Urbana

THE PERFORMANCE MONITOR FOR C.mmp (A MULTI-MINI-PROCESSOR 577
S. H. Fuller and D. K. Stevenson, Carnegie-Mellon University

IV-C: STOCHASTIC PROCESSES

Chairman: A. Haddad, University of Illinois, Urbana

ON HIGHER-ORDER SPECTRAL MOMENTS AND THE NORMALITY OF RANDOM PROCESSES: APPLICATION TO LINEAR AND NONLINEAR TRANSFORMATIONS 587
A. Traganitis and J. B. Thomas, Princeton University

ESTIMATION OF A RANDOM SPACE-TIME FIELD THAT INFLUENCES AN OBSERVED SPACE-TIME POINT PROCESS 597
P. M. Fishman and D. L. Snyder, Washingon University

STATISTICAL PROPERTIES OF THE SCATTERING CHANNEL 601
L. F. Eastwood, Jr., Washington University

ON A RELATIONSHIP BETWEEN LIKELIHOOD RATIOS AND CONDITIONAL MEAN ESTIMATES FOR THE EXPONENTIAL FAMILY OF DISTRIBUTIONS 605
S. C. Schwartz, Princeton University

CHARACTERIZATION OF NON-RENEWAL ERROR-PROCESSES 616
J.-P. A. Adoul, University of Sherbrooke, Canada

IV-D: NETWORKS II

Chairman: D. Calahan, University of Michigan

FACTORIZABILITY OF COEFFICIENTS OF AN LCR NETWORK DETERMINANT 627
Y. Kajitani, Tokyo Institute of Technology, Japan

EFFICIENT ANALYSIS OF NONLINEAR NETWORKS BY PIECEWISE-LINEAR APPROXIMATION 635
I. Hajj, K. Singhal and J. Vlach, University of Waterloo, Canada

EFFECT OF AUGMENTING REACTIVE ELEMENTS IN AN IMPROPER NETWORK ON ITS STATE EQUATIONS 643
M. A. Tapia, Georgia Institute of Technology

AN ALGORITHM FOR COMPUTING LAPLACE AND Z TRANSFORM OF A TIME SERIES WITH APPLICATION TO SYSTEM IDENTIFICATION 651
M. J. Corinthios, Ecole Polytechnique de Montreal, Canada

NOVEL APPLICATION OF THE NUMERICAL LAPLACE TRANSFORM INVERSION 661
M. Nakhla (Mrs.), K. Singhal and J. Vlach, University of Waterloo, Canada

FURTHER CONSIDERATIONS ON THE THEORY OF ORTHOGONAL KIRCHHOFF MODAL ANALYSIS 668
S. Shinoda, Chu-oh University, Japan and
Y. Kajitani, Tokyo Institute of Technology, Japan

V-A: SINGULAR PERTURBATIONS
Chairman and Organizer: P. Sannuti, Rutgers State University

EXAMPLES ILLUSTRATING THE CONNECTION BETWEEN SINGULAR PERTURBATIONS AND SINGULAR ARCS 678
R. E. O'Malley, Jr., University of Arizona

THE CONNECTION BETWEEN SINGULAR PERTURBATIONS AND SINGULAR ARCS: PART 2: A THEORY FOR THE LINEAR REGULATOR 686
A. Jameson, New York University

A CONCEPTUAL APPROACH TO APPLYING SINGULAR PERTURBATION METHODS TO VARIATIONAL PROBLEMS 693
A. J. Calise and R. Aggarwal, Dynamics Research Corporation, Wilmington, Massachusetts

ASYMPTOTIC STABILITY IN NETWORKS WITH PARASITICS 703
F. Hoppensteadt, Courant Institute of Mathematical Sciences, New York University

OPTIMIZATION OF A COUPLED-CORE NUCLEAR REACTOR SYSTEM BY THE METHOD OF ASYMPTOTIC EXPANSIONS 708
P. B. Reddy, Dynamics Research Corporation, Wilmington, Massachusetts and P. Sannuti, Rutgers University

APPLICATION OF SINGULAR PERTURBATION TO OPTIMAL ESTIMATION 718
H. E. Rauch, Lockheed Palo Alto Research Laboratory, Palo Alto, California

V-B: COMPUTER AND CIRCUIT DESIGN
Chairman: M. E. VanValkenburg, Princeton University

PARALLEL SOLUTION OF SPARSE SIMULTANEOUS LINEAR EQUATIONS 729
D. A. Calahan, The University of Michigan

A SYMBOLIC METHOD TO COMPUTER-AIDED DESIGN OF LINEAR ACTIVE NETWORKS BASED ON A NODAL APPROACH 739
N. B. Rabbat, University of Alabama and Washington University

MINIMAX OPTIMIZATION OF RECURSIVE DIGITAL FILTERS 748
C. Charalambous, University of Waterloo, Canada

A COMPUTER-AIDED SOLUTION TO THE LOW-PASS APPROXIMATION PROBLEM FOR DISTRIBUTED RC NETWORKS 758
R. I. Egbert and J. J. Bourquin, University of Missouri at Rolla

THE STEEPEST DESCENT METHOD FOR FINDING ROOTS OF POLYNOMIALS 760
J. R. Mitchell and W. L. McDaniel, Jr., Mississippi State University

V-C: DIGITAL SYSTEMS II
Chairman: D. Schertz, Bradley University

MINIMUM FEEDBACK SYNTHESIS WITH SMALLEST NUMBER OF STATE VARIABLES FOR SOC AND MOC ASYNCHRONOUS SEQUENTIAL MACHINES 763
D. Elhadef and E. J. Smith, Polytechnic Institute of Brooklyn

MODULAR DATA SWITCHING NETWORKS FOR MULTIPLE COMMUNICATION DEMANDS 771
G. M. Masson, University of Pittsburgh

CONTINUED FRACTIONS: A POWERFUL TOOL FOR DIGITAL ARITHMETICS 782
Bracha-Barak, A., University of Illinois, Urbana

COMPUTER SCIENCE EDUCATION: A DIFFERENT MACHINE MODEL OF EFFECTIVE COMPUTATION 786
M. E. Kaliski, Northeastern University and J. B. Geller, City College of New York

IMPROVING THE RELIABILITY OF FAULT-TOLERANT DIGITAL SYSTEMS BY USING MANY-VALUED LOGIC 787
E. G. DuCasse, Brooklyn College and G. A. Metze, University of Illinois, Urbana

V-D: CONTROL SYSTEMS II
Chairman: B. C. Kuo, University of Illinois, Urbana

THE DECENTRALIZED STABILIZATION AND CONTROL OF A CLASS OF UNKNOWN NON-LINEAR TIME-VARYING SYSTEMS 790
E. J. Davison, University of Toronto, Canada

STABILIZATION OF MULTI-INPUT CONTROLLABLE LINEAR SYSTEM WITH UNKNOWN SYSTEM PARAMETERS 798
C. Ho, United Technology Laboratory, Inc., Garland, Texas

DIGITAL COMPENSATOR SYNTHESIS 801
B. E. Bona, Rockwell International, Anaheim, California

LOW ORDER MODELS FOR DYNAMIC SYSTEMS 815
J. Cantalapeidra, Spanish Navy, N. V. Lam, South Vietnamese Navy and G. J. Thaler, Naval Postgraduate School, Monterey, California

AN EXPERIMENTAL STUDY OF A HYBRID ADAPTIVE CONTROL SYSTEM 816
E. F. Lizewski and R. V. Monopoli, University of Massachusetts

PARAMETER PLANE METHODS FOR LINEAR SYSTEMS WITH TIME DELAY 826
A. Hadjiconstandis, Hellenic Navy and G. J. Thaler, Naval Postgraduate School, Monterey, California

V-E: OTPIMIZATION
Chairman: K. Steiglitz, Princeton University

ON RENDEZVOUS WITH TERMINAL CONSTRAINT 836
Dong Hak Chyung, University of Iowa

APPLICATION OF THE EPSILON METHOD TO OPTIMAL CONTROL COMPUTATION 838
S. Mukherjee and N. K. Sinha, McMaster University, Canada

DISCRETE OPTIMIZATION 840
C. Charalambous, University of Waterloo, Canada

APPLICATION OF THE SYMMETRIC ALGORITHM TO SOLVE QUADRATIC PROGRAMMING PROBLEMS 848
J. B. Combot and S. H. Wu, Marquette University

SNAPSHOT SCHEDULING RULES FOR SERVICE SYSTEMS WITH NONLINEAR COSTS 850
P. B. Henderson and K. Steiglitz, Princeton University

OPTIMAL REPAIRMAN ASSIGNMENT 860
L. Shaw, Polytechnic Institute of New York

ON CANONICAL REALIZATIONS WITH UNBOUNDED INFINITESIMAL GENERATORS*

JOHN S. BARAS
Electrical Engineering Department
University of Maryland
College Park, Maryland 20742

Abstract

In this paper we study canonical (controllable and observable) realizations for infinite dimensional linear systems. In these realizations the infinitesimal state transition operator is unbounded but the generator of a C_0-semigroup in a Hilbert space. We present ways to reduce a realization to a canonical one. We also study the relation between the analytic properties of a transfer function and the spectral properties of the infinitesimal generators in its realizations. Finally we describe a class of transfer functions which can be realized by a system with infinitesimal generator having spectral properties closely related to the singularities of the transfer functions.

0. INTRODUCTION

The theory of infinite dimensional linear systems has progressed significantly in the last few years. This is mainly due to the application of certain fairly recent operator theoretic results in the study of infinite dimensional linear systems in Hilbert spaces. In particular the circle of ideas related to invariant subspace theory in H^2 spaces, provides a natural framework for a very successful study of the realization problem for non-rational transfer functions. Based on this approach several problems in infinite dimensional realization theory have been studied in our previous work [1] (with Roger W. Brockett), [2], by Paul A. Fuhrmann in [3], [4] and by J. W. Helton in [5].

The problem of realization, in our setting, is to express a given real function T defined on $[0,\infty)$ as $T(t) = c[e^{At}b]$, or to express its Laplace transform $\tilde{T}(s)$ as $c[(Is-A)^{-1}b]$ in some appropriately defined region of the complex plane. T is called the weighting pattern [6] and $\tilde{T}$ the transfer function. For further details, notation and relations to external and internal description of systems we refer to [1], [2]. In [1] we considered the following two cases: when A generates a C_0-semigroup on $\mathcal{H}$, $b \in \mathcal{H}$ and c is a bounded linear functional on $\mathcal{H}$, $[A, b, c]$ is a regular realization; in that case we write also $y(t) = \langle c, x(t) \rangle$ where c belongs to $\mathcal{H}$; $[A, b, c]$ is a balanced realization when A is as before, b belongs to the domain of A (denoted $\mathcal{D}_0(A)$), and c is a linear functional, defined on $\mathcal{D}_0(A)$ and such that $|c[x]| \le k\,(\|Ax\| + \|x\|)$ for $x \in \mathcal{D}_0(A)$ and some constant k. For a comparison between these two concepts we refer to

---

*This work was supported by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0298-0006, and represents a part of the author's Ph. D. dissertation, Harvard University, September 1973.

[1]. Our motivation for introducing the notion of a balanced realization comes from systems governed by partial differential equations with observations on the boundary of the domain of definition (see[2]p. 34, and [8] p. 200). For general motivation and examples we refer to [1], [2].

The following theorem ([1], Theorem 3) proves that the class of weighting patterns which admit balanced realizations is identical with the class of weighting patterns which admit regular realizations.

Theorem 1: A weighting pattern T has a balanced realization if and only if it has a regular one. Moreover the infinitesimal generators in both realizations can be taken to be the same.

We denote by Π^+ the half-plane Res>0. $H^2(\Pi^+)$ consists of functions analytic in Π^+ and square integrable along vertical lines in Π^+ such that

$$\sup_{x>o} \int_{-\infty}^{\infty} |\psi(x+iy)|^2 dy \leq M < \infty$$

We denote by $\mathbb{I}$ the imaginary axis in the complex plane. The Fourier transform

$$g(t) \xrightarrow{\mathfrak{F}} \int_{-\infty}^{\infty} e^{-i\omega t} g(t)dt = G(i\omega) \tag{1}$$

is a unitary map between $L_2(-\infty,\infty)$ and $L_2(\mathbb{I}, d\omega/2\pi)$. $L_2(o,\infty)$ is considered as the subspace of $L_2(-\infty,\infty)$, of functions which vanish on $(-\infty, o)$. $H^2(\mathbb{I}) = \mathfrak{F} L_2(o,\infty)$ and $\tilde{H}^2(\mathbb{I}) = \mathfrak{F} L_2(-\infty, o)$. We have also that $H^2(\mathbb{I})^{\perp} = \tilde{H}^2(\mathbb{I}) = \overline{H^2(\mathbb{I})}$. $H^2(\mathbb{I})$ consists exactly of the boundary values of the elements of $H^2(\Pi^+)$. The Paley-Wiener Theorem establishes that $H^2(\Pi^+) = \mathcal{L} L_2(o,\infty)$, where $\mathcal{L}$ denotes the Laplace transform. $\mathbb{D}$ denotes the open unit disk and $\mathbb{T}$ the unit circle. $H^2(\mathbb{T})$ denotes the subspace of $L^2(\mathbb{T})$ of functions with vanishing negative Fourier coefficients. $H^2(\mathbb{D})$ denotes the space of functions analytic in $\mathbb{D}$, with Taylor series around zero having square summable coefficients. $H^2(\Pi^+)$, $H^2(\mathbb{I})$, $\tilde{H}^2(\mathbb{I})$, $H^2(\mathbb{D})$ and $H^2(\mathbb{T})$ are called Hardy spaces.

In [1] we derived also criteria for a transfer function to admit a regular (and hence a balanced) realization. The most general conditions for realizability derived in [1] are given by the following theorem and corollary ([1] Theorem 6, Corollary 6.1).

Theorem 2: Let $T \in L_2(o,\infty)$ and continuous. If $\tilde{T}(i\omega) = F_c(i\omega) F_o(i\omega)$, where F_c, F_o belong to $H^2(\mathbb{I})$, then T has a regular realization.

Corollary 2.1: Let T be continuous and of exponential order. If for some α the function $T_1(t) = e^{-\alpha t} T(t)$ satisfies the conditions of Theorem 2, then T has a regular realization.

We denote, as usual, by e^{At} the C_0-semigroup of bounded operators generated by A. An element g of $\mathcal{H}$ is called a cyclic vector of the semigroup e^{At}, whenever the linear span of the vectors $e^{At}g$, $t \geq o$ is dense in $\mathcal{H}$. A realization [A, b, c] (balanced or regular) is controllable if and only if b is a cyclic vector of the semigroup e^{At} (Fattorini [9], p. 393). A regular realization [A, b, c] is observable if and only if c is a cyclic vector of the semigroup $e^{A^* t}$. A balanced realization is observable if and only if $x \in \mathcal{D}_0(A)$ and $c[e^{At}x] = 0$ for $t \geq 0$ imply $x = 0$. This implies that any two states that can be reached exactly are distinguishable. (For a summary of Controllability and Observability theory for infinite dimensional linear systems see [2]p. 59-65). A realization is canonical whenever is controllable and observable. In section 1 of this paper we show how to construct a canonical regular (or balanced) realization starting from a given regular (or balanced) realization.

It is often desirable that the internal model for a given input-output map satisfy certain requirements (besides simplicity) which are due to engineering considerations. The importance of the connectedness of the resolvent set of the infinitesimal generator in relation to frequency response methods for system identification is explained in [1]. The requirement that the spectral properties of the infinitesimal generator of a model should reflect closely the properties of the singularities of $\tilde{T}$ is essential from the engineering point of view. It is used as a guide in many ad hoc modeling or synthesis methods in electrical engineering [11], [12]. In section 3 of this paper we construct canonical realizations satisfying both these requirements. We characterize also a class of transfer functions which admit such realizations. It turns out that this class includes many of the transfer functions appearing in practical problems.

1. REDUCTION OF REALIZATIONS TO CANONICAL ONES

The following theorem describes a way to construct a canonical regular realization for a weighting pattern T, starting from any regular realization of T.

Note: When L is a closed subspace of $\mathcal{H}$ and B a linear operator, $P_L B|_L$ denotes the operator 'B restricted on L,' provided it is well defined.

Theorem 3: Let $[A, b, c]$ be a regular realization of a weighting pattern T on the Hilbert space $\mathcal{H}$. Let M be the closure of the linear span of the vectors $e^{At}b$ with $t \geq 0$, and P_M the associated orthogonal projection. Then i) $[P_M A|_M, b, P_M c]$ is a regular realization of T, with state space M.

Let now N be the closed linear span in M of the vectors $P_M e^{A^*t} P_M c$ with $t \geq 0$, and let P_N be the associated orthogonal projection ($P_N: \mathcal{H} \to N$). Then ii) $[P_N A|_N, P_N b, P_M c]$ is a canonical regular realization of T, with state space N.

Proof: Obviously M is invariant under e^{At} for $t \geq 0$. Hence

$$P_M e^{At} P_M = e^{At} P_M \tag{2}$$

Let $S(t) = P_M e^{At}|_M$. Then $S(t_1)S(t_2) = S(t_1+t_2)$ for $t_1, t_2 \geq 0$, $S(0)$=identity on M and $S(t)$ is strongly continuous on M. Hence $S(t)$, $t \geq 0$, is a C_o-semigroup on M. It is clear that the infinitesimal generator of $S(t)$ is the operator $P_M A|_M = A|_M$, which has domain dense in M. Then we write in our usual notation $S(t) = \exp(P_M A|_M t)$. Now

$$<P_M c, e^{(P_M A|_M)t} b> = < c, P_M e^{At} b> = < c, e^{At} b> = T(t)$$

Since M, $\mathcal{H}$ are Hilbert spaces we have that

$$(S(t))^* = (P_M e^{At}|_M)^* = P_M e^{A^*t}|_M = e^{(P_M A^*|_M)t}$$

is also a C_o-semigroup on M with infinitesimal generator $P_M A^*|_M$. Obviously N is invariant under $P_M e^{A^*t}|_M$ for $t \geq 0$ and so

$$P_N e^{A^*t}|_N = P_M e^{A^*t}|_N \overset{\text{def}}{=} S_1(t) \tag{3}$$

Based on (3) we prove as above that $S_1(t)$ is a C_o-semigroup on N, with infinitesimal generator the operator $P_N A^*|_N$. We write again in standard notation $P_N e^{A^*t}|_N = \exp(P_N A^*|_N t)$. We have also that

$$P_N e^{At} P_N = P_N e^{At} P_M \tag{4}$$

Hence $P_N e^{At}|_N = (S_1(t))^*$ is a C_o-semigroup on N with inf. gen. $P_N A|_N$. Now

$$<P_M c, e^{(P_N A|_N)t} P_N b> = <P_M c, P_N e^{At} P_M b> = <c, e^{At} b> = T(t)$$

Moreover if $x \in N$ and $<P_M c, \exp(P_N A|_N t)x> = 0$ for all $t \geq 0$, then it follows using (3) that $<P_M e^{A^*t} P_M c, x> = 0$ for $t \geq 0$, and so $x=0$ by the definition of N. If now $x \in N$ and $<\exp(P_N A|_N t) P_N b, x> = 0$ for all $t \geq 0$, then it follows from (4) that $<P_N e^{At} P_M b, x> = <e^{At} b, x> = 0$ for $t \geq 0$, and so $x=0$ by the definition of M. This proves (ii).

We can now use this Theorem and Theorem 1 to produce a canonical balanced realization starting from a given balanced realization. First we show that if the regular realization in Theorem 1 is canonical, then the balanced one is canonical too. The relations connecting the balanced realization $[A, b, c]$ to the regular realization $[A, b_1, c_1]$ ([1] Theorem 3) are given by $b = (\lambda I - A)^{-1} b_1$, $c[x] = <c_1, (\lambda I - A)x>$ with $\lambda > 1$ and in $\rho(A)$. Suppose $x \in \mathcal{D}_o(A)$ and $c[e^{At}x] = 0$ for $t \geq 0$. Then $<c_1, e^{At}(\lambda I - A)x> = 0$ for $t \geq 0$ and hence $(\lambda I - A)x = 0$. This implies $x=0$ since $\lambda \in \rho(A)$. Suppose now $<e^{At} b, x> = 0$ for $t \geq 0$. Then $<e^{At} b_1, (\lambda I - A^*)^{-1} x> = 0$ and so $(\lambda I - A^*)^{-1} x = 0$. This implies $x=0$ since $\lambda \in \rho(A^*)$. Therefore $[A, b, c]$ is canonical.

Hence we have the following:

Corollary 3.1: Given the balanced realization $[A, b, c]$ of T, we construct the associated regular one $[A, b_1, c_1]$ (using Theorem 1), which we reduce according to Theorem 3 to a canonical regular realization $[F, g_1, h_1]$. Finally we construct (using Theorem 1) the associated balanced realization $[F, g, h]$. Then $[F, g, h]$ is a canonical balanced realization of T.

Based on the above we can restrict our study to canonical regular realizations.

2. SPECTRAL CONSIDERATIONS AND THE HARDY CLASS ON HALF-PLANES

In the rest of this paper we investigate the problem of constructing realizations with infinitesimal generators having spectral properties closely related to the singularities of $\tilde{T}$. Following [1] we denote by $\sigma(\tilde{T})$ the set of nonanayticity of $\tilde{T}$ and by $\rho_o(A)$ the connected component of $\rho(A)$ which contains the half-plane $\mathrm{Re}\, s > \beta$. ($\mathrm{Re}\, s > \beta$ is the half-plane contained in $\rho(A)$, described in the Hille-Yosida Theorem [7]). We showed in [1] that if $[A, b, c]$ is a realization of T (balanced or regular) and if $\sigma_o(A)$ is the complement of $\rho_o(A)$ we must have the spectral inclusion property

$$\sigma(\tilde{T}) \subseteq \sigma_o(A)$$

Questions about the connectedness of the resolvent set of the infinit. generators, the simplicity of spectrum, and the relations to analytic properties of the transfer functions were for the first time raised and answered in [1], for bounded realizations (i.e. when the infinitesimal generators are bounded). In contrast with the finite dimensional theory the spectrum of the infinitesimal generator in a canonical regular realization is not uniquely determined by the transfer function (see [1], [2], [3], [4] for counterexamples with bounded realizations, which are a special class of regular realizations).

Here we present a similar investigation for transfer functions which admit regular realizations. In the rest of this paper we restrict to the study of the class described in Theorem 2. Extensions of our results to include the larger class described in Corollary 2.1 are easy.

To proceed we need some facts from the theory of $H^2(\Pi^+)$ functions. For details we refer to Hoffman [12], Duren [14]. A function Φ is *inner* if it is analytic in Π^+, with $|\Phi(s)| \leq 1$ for $s \in \Pi^+$, while $|\Phi(i\omega)| = 1$ a. e. A function $H \in H^2(\Pi^+)$ is *outer* if its boundary value on the imaginary axis is a cyclic vector for the semigroup ' multiplication by $e^{-i\omega t}$' on $H^2(\mathbb{I})$. Equivalently H is outer if its inverse Laplace transform h, has the property that its right translations form a dense set in $L_2(0, \infty)$. A subpsace of $L_2(0, \infty)$ which is invariant under right translations is mapped by the Fourier transform to a subspace of $H^2(\mathbb{I})$ which is invariant under 'multiplication by $e^{-i\omega t}$', $t \geq 0$. Following Lax [15], we call such a subspace of $H^2(\mathbb{I})$ a *right translation invariant subspace.* The orthogonal complement of a right translation invariant subspace is called a *left translation invariant subspace* and is invariant under ' multiplication by $e^{i\omega t}$, $t \geq 0$, followed by projection on $H^2(\mathbb{I})$' . The inverse Fourier transform of a left translation invariant subspace is a subspace of $L_2(0, \infty)$ which is invariant under left translations followed by restriction to $(0, \infty)$. Clearly we have corresponding facts for such subspaces in $H^2(\Pi^+)$.

For clarity we use the variable z for complex numbers in $\mathbb{D}$, while the variable s is used for complex numbers in Π^+. The map

$$z \xrightarrow{\eta} s = \frac{1+z}{1-z} \tag{5}$$

maps $\mathbb{D}$ onto Π^+. It is well known [12], [17], that the map $\mathcal{V}$ defined by

$$[\mathcal{V} F](z) = g(z) = \frac{2}{1-z} F\left(\frac{1+z}{1-z}\right) \tag{6}$$

is a unitary map from $H^2(\Pi^+)$ onto $H^2(\mathbb{D})$. $\mathcal{V}$ restricted to boundary values is a unitary map of $H^2(\mathbb{I})$ onto $H^2(\mathbb{T})$. The inverse of $\mathcal{V}$ is given by

$$[\mathcal{V}^{-1} g](s) = F(s) = \frac{1}{s+1} g\left(\frac{s-1}{s+1}\right) \tag{7}$$

Let as usual $\mathcal{H}$ denote a Hilbert space and e^{At} a C_0-semigroup of contractions on $\mathcal{H}$, with generator A. The operator $B = (A+I)(A-I)^{-1}$ is a contraction, called the *cogenerator* of e^{At} (see [16] p. 141). The relation between the semigroup and its cogenerator is the following:

$$\left.\begin{aligned} e^{At} &= e_t(B), \quad t \geq 0 \\ B &= \lim_{t \to 0+} \phi_t(e^{At}) \end{aligned}\right\} \tag{8}$$

where $e_t(\lambda) = \exp(t\,\lambda + t/\lambda - 1)$, $t \geq 0$, $\phi_t(\lambda) = \lambda - 1 + t/\lambda - 1 - t$. Using these relations it is easy to show that a vector $b \in \mathcal{H}$ is a cyclic vector for B if and only if it is a cyclic vector for e^{At}. Moreover B^* is the cogenerator of the adjoint semigroup $(e^{At})^*$ and the relations described above are valid if we replace B with B^* and e^{At} by $(e^{At})^*$ (see [16] p. 143).

The semigroup of contractions ' multiplication by $e^{-i\omega t}$' on $H^2(\mathbb{I})$ has as its cogenerator B, the operator ' multiplication by $i\omega - 1/i\omega + 1$ '. Then for $g \in H^2(\mathbb{D})$ we have

$$[\mathcal{V} B \mathcal{V}^{-1} g](z) = z g(z) = [Ug](z) \tag{9}$$

Hence B is unitarily equivalent to the forward shift on $H^2(\mathbb{T})$ (denoted as usual by U). The adjoint semigroup is ' multiplication by $e^{i\omega t}$ followed by projection on $H^2(\mathbb{I})$' , and its cogenerator is B^* and hence it is unitarily equivalent to the backward shift on $H^2(\mathbb{T})$. Clearly $F \in H^2(\mathbb{I})$ is a cyclic vector for B (resp. B^*) on $H^2(\mathbb{I})$ if and only if $\mathcal{V}$ F is a cyclic vector for the forward (resp. backward shift) on $H^2(\mathbb{T})$. Thus we have proved the following:

Lemma 1: A function $F \in H^2(\Pi^+)$ is outer if and only if $\mathcal{V} F \in H^2(\mathbb{D})$ is outer. A function $F \in H^2(\mathbb{I})$ is a cyclic vector for the semigroup "multiplication by $e^{i\omega t}$ followed by projection on $H^2(\mathbb{I})$' if and only if $\mathcal{V} F \in H^2(\mathbb{T})$ is a cyclic vector for the backward shift on $H^2(\mathbb{T})$.

The following theorem is due to Lax and characterizes right translation invariant subspaces in $H^2(\Pi^+)$.

Theorem 4 ([15]): Every closed subspace R of $H^2(\Pi^+)$ invariant under ' multiplication by e^{-st}' , $t \geq 0$, is of the form $\Phi H^2(\Pi^+)$ where Φ is inner. Φ is unique modulo a constant of modulus one. If ϕ is the inner function associated to $\mathcal{V}$R by Beurling' s theorem then $\Phi(s) = \phi(\frac{s-1}{s+1})$.

An inner function Φ is normalized whenever the corresponding ϕ on the disk is normalized. Any element $F \in H^2(\Pi^+)$ has a factorization $F = \Phi \cdot H$ where Φ is inner and H is outer. Every inner function has a factorization $\Phi = cBS$ where c is a constant of modulus 1, B a Blaschke product and S a singular function (see [12]for details).

Using Lemma 1 we can obtain many properties of cyclic or non-cyclic vectors for the semigroup ' multiplication by $e^{i\omega t}$ followed by projection on $H^2(\mathbb{I})$' , from properties of cyclic or noncyclic vectors for the backward shift on $H^2(\mathbb{T})$ (see [13]). For example any $F \in H^2(\Pi^+)$ with isolated branch points on the imaginary axis is a cyclic vector.

In this paper cyclic or noncyclic is understood with respect to the semigroup ' multiplication by $e^{i\omega t}$ followed by projection on $H^2(\mathbb{I})$' .

Definition: Let $F \in H^2(\mathbb{I})$, e^{At} denote the semigroup ' multiplication by $e^{i\omega t}$ followed by projection on $H^2(\mathbb{I})$' . The left translation invariant subspace generated by F is the closure of the linear span of the vectors $e^{At}F$, $t \geq 0$, in $H^2(\mathbb{I})$.

Theorem 5: An element F of $H^2(\mathbb{I})$ is a noncyclic vector if and only if there exist an inner function Φ and a function H in $H^2(\Pi^+)$ such that $F(i\omega) = \overline{H(i\omega)}\,\Phi(i\omega)$ a. e. on $\mathbb{I}$. Moreover if we choose Φ to be normalized and relatively prime to the inner factor of H this factorization is unique. In this case the left translation invariant subspace generated by F is $(\Phi H^2(\mathbb{I}))^{\perp}$.

Proof: This is a direct consequence of the corresponding theorem on the disk ([13]p. 56) and of the properties of the map $\mathcal{V}$ defined by (6), (7).

Definition: The inner function uniquely associated to every non-cyclic vector $F \in H^2(\Pi^+)$ by Theorem 5 is called the associated inner function of F.

We are ready now to proceed with the study of canonical realizations for the class of weighting patterns described in Theorem 2.

For any such transfer function we have the 'right translation realization' which is constructed by considering as Hilbert space $\mathcal{H}$ the space $H^2(\mathbb{I})$,the semigroup ' multiplication by $e^{-i\omega t}$ ' as e^{At}, F_c as b and F_o as c. Using inverse Fourier transform this realization is described via the semigroup of right translations on $L_2(0,\infty)$. Our plan is to apply Theorem 3 to this realization, obtain a canonical one, and then discuss the spectral properties of the latter . We assume with no loss of generality that the inner factors of F_c and F_o are relatively prime. We start with the following preliminary result.

Lemma 2: Suppose $\tilde{T} \in H^2(\Pi^+)$ and has a factorization $\tilde{T}(i\omega) = \overline{F_c(i\omega)}\, F_o(i\omega)$ a. e. on the imaginary axis, where $F_c, F_o \in H^2(\Pi^+)$ and have no common inner factor. Then $\tilde{T}$ is noncyclic if and only if F_o is.

Proof: Suppose F_o is noncyclic. Then by Theorem 5, $F_o(i\omega) = \overline{H_o(i\omega)}\, \Phi_o(i\omega)$ a. e. on $\mathbb{I}$, where $H_o \in H^2(\Pi^+)$ and Φ_o is inner. Hence $\tilde{T} = \overline{F_c(i\omega) H_o(i\omega)}\, \Phi_o(i\omega)$ a. e. on $\mathbb{I}$. Since $\tilde{T} \in H^2(\Pi^+)$ we have that $F_c H_o \in H^2(\Pi^+)$ and therefore by Theorem 5, $\tilde{T}$ is noncyclic.

Suppose now that $\tilde{T}$ is noncyclic. Then by (6) we have

$$[\mathcal{V}\tilde{T}](e^{i\theta}) = \frac{2}{1-e^{i\theta}} \overline{F_c\left(\frac{1+e^{i\theta}}{1-e^{i\theta}}\right)}\, F_o\left(\frac{1+e^{i\theta}}{1-e^{i\theta}}\right) = \overline{\frac{1}{2} e^{i\theta} [\mathcal{V}F_c](e^{i\theta})}\,(e^{i\theta}-1)[\mathcal{V}F_o](e^{i\theta}) \tag{10}$$

Let now

$$\mathcal{V}\tilde{T} = g;\ \frac{1}{2} e^{i\theta}[\mathcal{V}F_c](e^{i\theta}) = f(e^{i\theta});\ \ (e^{i\theta}-1)[\mathcal{V}F_o](e^{i\theta}) = h(e^{i\theta}) \tag{11}$$

Then $g \in H^2(\mathbb{D})$ and by Lemma 1 is a noncyclic vector for the backward shift; $f \in H^2(\mathbb{D})$ and $f(0) = 0$; $h \in H^2(\mathbb{D})$. By Theorem 3.1.5 in [13], there exists an element $g_1 \in H^2(\mathbb{D})$ with $g_1(0) = 0$ and an inner function ϕ (on the disk) such that $g = \overline{g}_1 \phi$ a. e. on $\mathbb{T}$. Therefore

$$g = \overline{f} h = \overline{g}_1 \phi \tag{12}$$

Now $\phi f \in H^2(\mathbb{D})$. Let U denote, as usual, the forward shift and U* the backward shift on $H^2(\mathbb{D})$. Then

$$<\phi f, U^{*n} h> = <U^n \phi f, h> = \int \overline{z^n \phi f}\, h\, d\mu(\theta) = \int \overline{z^n}\, \overline{g}_1\, d\mu(\theta) = 0;\ n = 0, 1, 2, \ldots$$

because $\overline{g}_1 \perp H^2(\mathbb{D})$. Since $\phi f \perp U^{*n} h$, for $n = 0, 1, 2, \ldots$ it follows that h is noncyclic. But $v \in H^2(\mathbb{T})$ is a noncyclic vector for the backward shift if and only if $e^{i\theta} v$ is one (see [13] Theorem 2.2.8). We conclude therefore from (11) that $\mathcal{V} F_o$ is noncyclic. Then by Lemma 1, F_o is noncyclic and this completes the proof.

Since the properties of the 'right translation realization' depend heavily on $\tilde{T}$ being cyclic or noncyclic, we study these two cases separately, in the sequel.

3. NONCYCLIC TRANSFER FUNCTIONS

Since $\tilde{T}$ is noncyclic it follows from Lemma 2 that F_o is noncyclic.

We study first transfer functions, for which F_c is outer. Then the 'right translation realization' is controllable. Hence applying Theorem 3 we obtain a canonical realization with state space N = left translation invariant subspace generated by F_o in $H^2(\mathbb{I})$, with $P_N e^{-i\omega t}|_N$ as e^{At}, with F_o as c and $P_N F_c$ as b. Moreover by Theorem 5 since F_o is noncyclic we have that $N = (\Phi_o H^2(\mathbb{I}))^{\perp}$, where Φ_o is the associated inner function of F_o.

The following theorem describes the spectrum of the infinitesimal generator of the semigroup $P_N e^{-i\omega t}|_N$ (where $N = (\Phi_o H^2(\mathbb{I}))^{\perp}$) in terms of the inner function Φ_o.

Theorem 6 ([8] p. 70): Let N be a left translation invariant subspace of $H^2(\mathbb{I})$, i. e. $N = (\Phi H^2(\mathbb{I}))^{\perp}$ for some inner function Φ. Consider the semigroup 'multiplication by $e^{-i\omega t}$', restricted on N (i. e. $P_N e^{-i\omega t}|_N$). The spectrum of its infinitesimal generator is the set $\mathfrak{S}_\Phi$ which consists of

i) all complex numbers μ with Re $\mu < 0$, such that $\Phi(-\mu)=0$

ii) all complex numbers μ with Re $\mu = 0$ such that Φ cannot be continued analytically across the imaginary axis at $-\mu$.

Combining this result with the previous discussion we have

Theorem 7: Let $\tilde{T}$ be a transfer function which belongs to $H^2(\Pi^+)$, is noncyclic and has a factorization $\tilde{T} = F_c F_o$ a.e. on the imaginary axis with F_c, F_o in $H^2(\Pi^+)$ and F_c outer. Then $\tilde{T}$ has a canonical realization with the spectrum of the infinitesimal generator being exactly $\mathcal{S}_{\Phi_o}$, where Φ_o is the associated inner function of F_o. This realization is constructed by taking, as state space the subspace of $H^2(\mathbb{I})$, $N=(\Phi_o H^2(\mathbb{I}))^{\perp}$, as c the function F_o, as b the projection of F_c on N and as semigroup the restriction of the semigroup ' multiplication by $e^{-i\omega t}$' on N (i.e. $P_N e^{-i\omega t}|_N$).

Suppose now that $\tilde{T}$ has a meromorphic continuation across the imaginary axis in Π^- (i.e. it has a finite number of poles in any finite region of Π^-, which accumulate only on $\mathbb{I}$). Now

$$\tilde{T}(i\omega) = \overline{F_c(i\omega)}\, F_o(i\omega) = \overline{F_c(i\omega)H_o(i\omega)}\, \Phi_o(i\omega) \tag{13}$$

where H_o, Φ_o are the factors of F_o according to Theorem 5. Now since T is real valued we have

$$\tilde{T}(i\omega) = \bar{\tilde{T}}(-i\omega) = \frac{F_c(-i\omega)H_o(-i\omega)}{\Phi_o(-i\omega)} \tag{14}$$

The right hand side of (14) is the boundary value of the function

$G(s) = \dfrac{F_c(-s)H_o(-s)}{\Phi_o(-s)}$, which is meromorphic in Π^-. Since F_c does not have any inner factor and Φ_o is relatively prime to the inner factor of H_o we see that G is analytic in Π^- except at points $\mu \in \Pi^-$ where $\Phi_o(-\mu)=0$. Since $\tilde{T}$ has a mermorphic continuation in Π^- we must have that $\tilde{T}(s)=G(s)$ in Π^-. Hence the singularities of $\tilde{T}$ in Π^- are the points μ with $\Phi_o(-\mu)=0$.

Since T is real valued we have $\bar{\tilde{T}}(-\bar{s}) = \tilde{T}(-s)$. Therefore $\tilde{T}$ has an analytic continuation through $i\omega$ in Π^-, if and only if it has one through $-i\omega$. From (13) $\tilde{T}(i\omega) = \overline{F_c(i\omega)H_o(i\omega)}\,\Phi_o(i\omega)$, where Φ_o and the inner factor of $F_c H_o \in H^2(\Pi^+)$ are relatively prime. Hence by a theorem in [8]p. 66, $\tilde{T}$ has an analytic continuation in Π^- through $-i\omega$, if and only if Φ_o has one through $-i\omega$. Comparing $\sigma(\tilde{T})$ as described above with the set $\mathcal{S}_{\Phi_o}$ of Theorem 6, we see that for the canonical realization given by Theorem 7 the spectral inclusion property becomes again equality

$$\sigma(\tilde{T}) = \sigma(A)$$

This motivates the following definition.

Definition: A canonical regular (resp. balanced) realization [A, b, c] of a weighting pattern T is called S-minimal regular (resp. balanced) (S from spectrum) if and only if $\sigma(\tilde{T}) = \sigma(A)$, multiplicities counted whenever possible.

We have thus proved the following.

Corollary 7.1: Any transfer function which satisfies the conditions of Theorem 7 above and is meromorphic in the left half plane, has an S-minimal regular realization with infinitesimal generator having connected resolvent set. The construction of this realization is given by Theorem 7 above.

We discuss briefly an application of these results to a particularly interesting class of weighting patterns which belong to the class we are considering here. Recall (see [1]) that if T belongs to $L_2(0,\infty)$, is locally absolutely continuous, its derivative belongs to $L_2(0,\infty)$ and $T(0)=0$, then T belongs to the class we are studying. In this case we have $F_c(i\omega)=1/(1+i\omega)$, $F_o(i\omega)=(1-i\omega)\tilde{T}(i\omega)$. Now $[\mathscr{T}F_c](z)=1$ and by Lemma 1, F_c is outer. We thus have

Corollary 7.2: Let T be a weighting pattern which belongs to $L_2(0,\infty)$, is locally absolutely continuous, its derivative belongs to $L_2(0,\infty)$ and $T(0)=0$. If $\tilde{T}$ is noncyclic and meromorphic in the left half-plane, T has an S-minimal realization with infinitesimal generator having connected resolvent set. The construction is given in Theorem 7.

Remark: The meromorphic assumption is satisfied by many systems governed by several forms of the wave equation (e.g. Schrödinger Equation, Maxwell's Equations) (see [8] Ch. VI and appendix 4, [10]).

If F_c is not outer the 'right translation realization' is neither controllable nor observable. By Th. 3 we obtain first a realization with state space $M=\Phi_c H^2(\mathbb{I})$ where Φ_c is the normalized inner factor of F_c, with $e^{-i\omega t}|_M$ as e^{At}, $P_M F_o$ as c and F_c as b. Next we reduce the latter realization to obtain a canonical one. This has as state space N, the closure of the linear span of the vectors $P_M e^{i\omega t} P_M F_o$, $t\geq 0$, in M, $P_M F_o$ as c and $P_N F_c$ as b. The semigroup is $P_N e^{-i\omega t}|_N$. Since M is invariant under $e^{-i\omega t}$, $P_M e^{i\omega t} P_M F_o = P_M e^{i\omega t} F_o$ for $t\geq 0$. Since F_o is noncyclic, the left translation invariant subspace generated by F_o is $(\Phi_o H^2(\mathbb{I}))^{\perp}$ where Φ_o is the assoc. inner function of F_o. The situation here looks similar to the previous case. However the existence of the projection P_M complicates the discussion of the spectral properties of this realization.

4. CYCLIC TRANSFER FUNCTIONS

Since $\tilde{T}$ is cyclic, it follows from Lemma 2 that F_o is cyclic also. This class is very interesting because it contains transfer functions with branch points, such as those usually appearing in systems governed by partial differential equations. By the spectral inclusion property all the points on branch cuts of $\tilde{T}$ are included in the spectrum of any infinitesimal generator with connected resolvent set which realizes $\tilde{T}$. Hence there is no unique "minimal" spectrum, due to the nonuniqueness of the branch cuts. When F_c is outer, the 'right translation realization' is canonical. The spectrum of the infinitesimal generator in this realization is however the whole closed left half-plane and hence generally it is far from being equal to $\sigma(\tilde{T})$. So again (see [1]) canonical by no means implies S-minimal. When F_c is not outer the 'right translation realization' is observable. Reducing this realization by Theorem 3 we obtain a canonical realization with state space $M=\Phi_c H^2(\mathbb{I})$ where Φ_c is the normalized inner factor of F_c, with $P_M e^{-i\omega t}|_M$ as e^{At}, with $P_M F_o$ as c and F_c as b. Since M is invariant under 'multiplication by $e^{-i\omega t}$' it is easy to see that the spectrum of the infinitesimal generator in this realization is the whole closed left half-plane, and so this realization is generally far from being S-minimal.

Therefore independently of F_c being outer or not, whenever $\tilde{T}$ is cyclic the 'right translation realization' does not reduce to an S-minimal realization in general. Only when $\tilde{T}$ is noncyclic, we can reduce the 'right translation realization' to obtain an S-minimal one.

It may be possible to construct S-minimal realizations for $\tilde{T}$ cyclic, by other means however (compare with [1], section 6).

We conclude this paper with the following.

Remark: We do not have a complete picture for the relations between canonical (resp. S-minimal) realizations of the same weighting pattern T, for the class considered in this paper. It is apparent however that there is no analogue of the state space isomorphism theorem of the finite dimensional theory. Counterexamples similar to the ones presented for the bounded case (see [1],[2],[3],[4]), can be constructed easily. To obtain a state space isomorphism theorem we need more assumptions. J. W. Helton [5] introduced the notions of exact controllability and observability and obtained a state space isomorphism theorem for bounded realizations. This result can be extended in a straightforward manner to regular and balanced realizations.

Acknowledgements: I would like to thank Roger Brockett and Paul Fuhrmann for many stimulating discussions during the course of this work.

REFERENCES

1. J. S. Baras and R. W. Brockett, "H^2 Functions and Infinite Dimensional Realization Theory," submitted to SIAM J. on Control. An earlier version appeared in the Proc. of the IEEE Decision and Control Conf., 1972.
2. J. S. Baras, Intrinsic Models for Infinite Dimensional Linear Systems, Ph. D. thesis, Harvard University, Sept. 1973.
3. P. A. Fuhrmann, "On Realization of Linear Systems and Applications to Some Questions of Stability," to appear Math. Sys. The.
4. P. A. Fuhrmann, "Exact Controllability and Observability and Realization Theory in Hilbert Space," to appear.
5. J. W. Helton, "Discrete Time Systems, Operator Models and Scattering Theory," to appear.
6. R. W. Brockett, Finite Dimensional Linear Systems, John Wiley, 1970.
7. A. Friedman, Partial Differential Equations, Holt, Rinehart and Winston 1969.
8. P. D. Lax and R. S. Phillips, Scattering Theory, Academic Press, 1967.
9. H. O. Fattorini, "On Complete Controllability of Linear Systems," Journal of Differential Equations, 3, 391-402, 1967.
10. H. J. Carlin and D. Noble, "Circuit Properties of Coupled Dispersive Lines with Applications to Wave Guide Modeling," to appear.
11. R. Levy "A Generalized Design Technique for Practical Distributed Reciprocal Ladder Networks," IEEE Trans. Microwave Theory Tech., Vol. MTT-21, No. 8, p. 519-526, August 1973.
12. K. Hoffman, Banach Spaces of Analytic Functions, Prentice Hall, 1962.
13. R. G. Douglas, H. S. Shapiro, A. L. Shields, "Cyclic Vectors and Invariant Subspaces for the Backward Shift Operator", Ann. Inst. Fourier, Grenoble 20, 1, 1971, 37-76.
14. P. L. Duren, Theory of H^p Spaces, Academic Press, 1970.
15. P. D. Lax, "Translation Invariant Spaces," Acta. Math., Vol. 101, 1959.
16. Be' la Sz-Nagy and C. Foias, "Harmonic Analysis of Operators on Hilbert Space," N. Holland and Am. Elsevier, 1970.
17. J. W. Moeller, "On the Spectra of Some Translation Invariant Spaces," J. Math. Anal. Appl. 4, 1962, 276-296.

THE REGULATOR PROBLEM AND STABILIZATION OF INFINITE DIMENSIONAL CONTROL SYSTEMS

RICHARD DATKO
Georgetown University, Washington, D. C. 20007

Abstract

In this report some aspects of the regulator problem for control problems in infinite dimensional settings shall be discussed. Because of space limitations no examples will be given. However, examples of the problems mentioned can be found in [6], [9], [11] or [12]. Two general classes of problems will be discussed. The first concerns what are called linear evolutionary processes in a Banach space and the second linear functional differential equations of neutral or retarded type. For simplicity of presentation and because they are the most frequently used spaces we shall limit ourselves in Section 1 to real Hilbert spaces.

1. THE REGULATOR PROBLEM FOR REAL HILBERT SPACES

Assumptions and Notation. H_i, $i = 1, 2$ will denote real Hilbert spaces with inner product $(\cdot,\cdot)$. $L_2(H_2)$ will denote the equivalence classes of measurable mappings from $[0,\infty) \to H_2$ whose norms are square integrable. $S(t,s)$ will denote a family of bounded linear mappings of H_1 into itself which are strongly continuous on the infinite triangle $\Delta = \{(t,s): 0 \leq s \leq t < \infty\}$. The family $S(t,s)$ will satisfy the conditions (i) $S(t,t) = I$ (the identity mapping) for all t $[0,\infty)$, (ii) if $t_o \leq s \leq t$ then $S(t,s)S(s,t_o) = S(t,t_o)$ and (iii) there exist constants $a \geq 0$ and $M \geq 1$ such that for all $(t,t_o) \varepsilon \Delta$ the norm $|S(t,t_o)|$ satisfies $|S(t,t_o)| \leq Me^{a(t-t_o)}$. The adjoint family, $S^*(t,s)$, to $S(t,s)$ will also be strongly continuous on Δ. $B(t)$ will be a strongly continuous and uniformly bounded mapping from H_2 into H_1; its adjoint, $B^*(t)$, will also be strongly continuous. $W(t)$ will be a strongly continuous symmetric positive mapping from H_1 into itself in the sense that for all x,y in H_1 and $t \varepsilon [0,\infty)$ there exists a constant $m > 0$ such that $(W(t)x,y) = (W(t)y,x)$ and $0 < (W(t)x,x) \leq m(x,x)$. $U(t)$ will be a strongly continuous symmetric positive definite mapping from H_2 into itself in the sense that for all u,v in H_2 and $t \varepsilon [0,\infty)$ there exists positive constants m_1 and m_2 such that $(U(t)u,v) = (U(t)v,u)$ and $m_1(u,u) \leq (U(t)u,u) \leq m_2(u,u)$. This last condition ensures the existence of $U^{-1}(t)$ as a strongly continuous mapping.

We wish to minimize the cost functional

$$C(u,x_o,t_o,T) = \int_t^T [(W(t)x_u(t,t_o,x_o),x_u(t,t_o,x_o))+(U(t)u(t),u(t))]dt \quad (1)$$

subject to the constraint

$$x_u(t,t_o,x_o) = S(t,t_o)x_o + \int_{t_o}^T S(t,s)B(s)u(s)ds \ . \quad (2)$$

Here we assume $0 \leq t_o \leq T \leq \infty$, $x_o \in H_1$ and $u \in L_2(H_2)$. We shall first consider the case $T < \infty$ and obtain results for $T = \infty$ as a limiting case.

Using the constraint (2) the Fréchet derivative of the functional $C(\cdot,x_o,t_o,T)$ can be computed and satisfies the equation

$$C'(u,x_o,t_o,T) = 2[U(t)u(t)+B^*(t)\int_t^T S^*(s,t)W(s)x_u(s,t_o,x_o)ds] \quad (3)$$

Placing this derivative equal to the zero vector in $L_2(H_2)$ we obtain the expression

$$u(t) = -U^{-1}(t)B^*(t)\int_t^T S^*(s,t)W(s)x_u(s,t_o,x_o)ds \ . \quad (4)$$

It is shown in [7] that the optimal control (i.e., minimizing u) and the optimal trajectory (i.e., the solution of (2) corresponding to the optimal control) exist and are unique and uniquely satisfy (4). We denote this pair by u^m and x^m.

For each pair $0 \leq t \leq T$ we define the mapping $K(T,t)$ from $H_1 \to H_1$ by the relation

$$K(T,t)x_o = \int_t^T S^*(s,t)W(s)x^m(s,t_o,x_o)ds \ . \quad (5)$$

It has been shown in [7] that $K(T,t)$ is a continuous linear mapping which satisfies $(K(T,t)x,y) = (K(T,t)y,x)$, $(K(T,t_o)x_o,x_o) = C(u^m,x_o,t_o,T)$ and $K(T,t)x_o = K(T,t)x^m(t,t_o,x_o)$ for each $t \in [t_o,T]$. Using these properties we can prove the theorem (see e.g. [7]):

Theorem 1. If $t_o \leq t < T < \infty$, then the optimal control and optimal trajectory are unique and the optimal control satisfies the relation

$$u^m(t,t_o,x_o) = -U^{-1}(t)B^*(t)K(T,t)x^m(t,t_o,x_o) \ . \quad (6)$$

Equation (6) is a feedback characterization of the optimal control for finite time intervals. We would like to obtain an extension of this result to infinite intervals. This cannot be done without additional assumptions. Thus we make the following hypothesis.

Hypothesis 1. For every x_o, independently of the interval $[t_o,T]$, there exists a real number $M(x_o)$ such that $C(u,x_o,t_o,T) \leq M(x_o)$ for some $u \in L_2(H_2)$.

Conditions under which Hypothesis 1 holds are not always easy to obtain. For finite dimensional systems a sufficient condition for Hypothesis 1 to hold is for the system to be completely controllable (see e.g. [1]), with the controllability being uniform over intervals of the form $[t_o,t_o+\varepsilon_o]$ where $\varepsilon_o > 0$ is fixed. Another sufficient condition is that the uncontrolled system is uniformly asymptotically stable (see e.g. [5]). For finite dimensional autonomous systems there are intermediate conditions between the above two which are sufficient for Hypothesis 1 to hold. Suffice it to say that in [7] conditions are given which are sufficient for Hypothesis 1. However, we shall omit their explicit statement here. The next theorem is proved in [7].

Theorem 2. If Hypothesis 1 holds there exists a family of linear mappings, $K(t)$, from H_1 into itself such that for any x,y in H_1 and $t \in [0,\infty)$ $(K(t)x,y) = (K(t)y,x)$ and $(K(t)x,x) = C(u^m,x,t,\infty)$. For x and t fixed $\lim_{T\to\infty} K(T,t)x = K(t)x$ and for each $x \in H_1$ the function $K(t)x$ is continuous on $[0,\infty)$.

Corollary. If $U(t)$, $W(t)$ and $B(t)$ are constant mappings and $S(t,t_o) = S(t-t_o,0)$ for all $(t,t_o) \in \Delta$ then the mapping $K(t)$ of Theorem 1 is constant.

The next theorem which is proved in [7] characterizes the optimal controls and optimal trajectories over an infinite interval.

Theorem 3. If Hypothesis 1 holds, then the optimal control over an infinite interval is unique and given by the expression

$$u^m(t) = -U^{-1}(t)B^*(t)K(t)x^m(t,t_o,x_o) . \tag{7}$$

There exist only one pair of functions $(u^m(t),x^m(t))$ defined from $[t_o,\infty) \to H_2 x H_1$ which satisfy equation (7). The collection of optimal trajectories generates a new linear evolutionary process, $T(t,t_o)$, defined on Δ which satisfies an estimate on its norm of the form $|T(t,t_o)| \leq Me^{\sigma(t-t_o)}$ where $M \geq 1$ and $-\infty < \sigma < \infty$. An explicit integral characterization of this process is given by

$$T(t,t_o)x_o = S(t,t_o)x_o - \int_{t_o}^{t} S(t,s)B(s)U^{-1}(s)B^*(s)K(s)T(s,t_o)x_o ds. \quad (8)$$

A natural question to ask is how the results stated in Theorem 3 specialize to finite dimensional systems. This is answered by the following corollary.

Corollary. If in Theorem 3 the control problem (1) - (2) is finite dimensional and $S(t,t_o)$ satisfies the equation $\frac{d}{dt}[S(t,t_o)] = A(t)S(t,t_o)$ where $A(t)$ is an nxn matrix, then the optimal trajectories are given by solutions to the differential equation $\frac{d}{dt}(x(t)) = [A(t)-B(t)U^{-1}(t)B^*(t)K(t)]x(t)$. The matrix $K(t)$ satisfies the matrix differential equation

$$\frac{dK}{dt} + K(t)A(t) + A^*(t)K(t) - K(t)B(t)U^{-1}(t)B^*(t)K(t) = -W(t) .$$

In the finite dimensional case it is known (see e.g. [1]) that if the system is autonomous, completely controllable and the matrix W is positive definite then the feedback control generates an exponentially stable system. Surprisingly enough a similar result holds for nonautonomous systems without regard to their dimension.

Theorem 4. If Hypothesis 1 holds and $W(t)$ is positive definite in the sense that there exists constants $0 < w_1 < w_2$ such that $w_1(x,x) \leq (W(t)x,x) \leq w_2(x,x)$ for all $x \in H_1$, then the optimal trajectories generate an exponentially stable evolutionary process, i.e. $|T(t,t_o)| \leq Me^{\sigma(t-t_o)}$ where $\sigma < 0$.

2. THE REGULATOR PROBLEM FOR FUNCTIONAL EQUATIONS

In this section we shall discuss the regulator problem for retarded and neutral functional differential equations. All operators in this section are assumed to be continuous and uniformly bounded matrices of the appropriate dimensions. The phase space H_1 is now Euclidean R^n and the control space H_2 is Euclidean R^m. The problem is to minimize the quadratic functional on $L_2(R^m)$ given by

$$C(u,\phi,t_o,T) = \int_{t_o}^{T} [(W(t)x_u(t,t_o,\phi),x_u(t,t_o,\phi)) + (U(t)u(t),u(t))]dt, \quad (9)$$

where for $0 \leq t_o \leq t$ $x_u(t,t_o,\phi)$ is an n-vector satisfying the functional differential equation

$$\frac{d}{dt}[x(t) - \sum_{i=1}^{N} A_i x(t-h_i)] = \sum_{i=0}^{N} C_i(t)x(t-h_i) + D(t)u(t). \quad (10)$$

In (10) $0 = h_o < h_1 < \ldots < h_n$ and if $t \varepsilon [t_o-h_n,t_o]$ then $x_u(t,t_o,\phi) = \phi(t)$ a continuous n-vector. The space of all continuous n-vectors defined on $[t_o-h_n,t_o]$ will be denoted by C. C can be equipped with a norm which makes it a Banach space; the topological dual of C is the space of n-vectors of bounded variation, BV. If $g \varepsilon BV$ then the continuous linear functional induced by g on C will be denoted by $<g,\cdot>$ (see e.g. [8]).

We observe that in (10) the matrices A_i, $1 \leqslant i \leqslant N$, are constant. Thus the results presented here are not for a completely nonautonomous system. We shall also assume $\det [I - \sum_{i=1}^{N} A_i e^{-\lambda h_i}] = 0$ either has no solutions in the complex plane or all solutions lie in a left half plane $Re\lambda \leqslant -\sigma, \sigma > 0$. In the former case the control problem is said to be retarded in the later it is said to be neutral.

In conjunction with the system (10) we shall consider its "adjoint". This is the system

$$\frac{d}{dt}[y(t)-\sum_{i=1}^{N} A_i^* y(t+h_i)] = -\sum_{i=0}^{N} C_i^*(t+h_i)y(t+h_i) + g(t) . \qquad (11)$$

It is shown in [3] that all solutions of (10) can be represented by the variation of parameters formula

$$x_u(t,t_o,\phi) = z(t,t_o,\phi) + \int_{t_o}^{t} X(t,s)D(s)u(s)ds, \qquad (12)$$

where $z(t,t_o,\phi)$ is the solution of the homogeneous part of (10), i.e. when $u(t) \equiv 0$, and $S(t,s)$ is a matrix solution in s of (11) for $g(t) \equiv 0$ which also satisfies the conditions $S(t,t) = I$, $S(t,s) = 0$ for $s > t$ and the matrix $S(t,s) - \sum_{i=1}^{N} A_i^* S(t,s+h_i)$ is continuous in t for $t \geqslant s$.

Let us define the continuous n-vectors

$$x_t^{\,u}(\phi) = \{x_u(s,t_o,\phi) : t-h_n \leqslant s \leqslant t\}$$

and

$$y_t = \{y(s) : y(s) \text{ satisfies (11) on } [t,t-h_n]\} . \qquad (13)$$

If y_t is differentiable almost everywhere on $[t,t+h_n]$ with an integrable derivative we define the bilinear form

$$\langle y_t, x_t^u \rangle = (y(t), x_u(t) - \sum_{i=1}^{N} A_i x(t-hi)) +$$

$$\sum_{i=1}^{N} \int_{t-h_i}^{t} (y(\alpha+h_i), C_i(\alpha+h_i)x_u(\alpha))d\alpha -$$

$$\sum_{i=1}^{N} \int_{t-h_i}^{t} (\frac{d}{d\alpha}y(\alpha+h_i), A_i x(\alpha))d\alpha \ . \tag{14}$$

A brief computation yields the relationship

$$\langle y_t, x_t^u \rangle = \langle y_{t_o}, x_{t_o}^u \rangle + \int_{t_o}^{t} (y(\alpha), D(\alpha)u(\alpha))d\alpha + \int_{t_o}^{t} (g(\alpha), x_u(\alpha))d\alpha \ . \tag{15}$$

If $T < \infty$ then exactly as in Section 1 we compute the derivative of $C(u,\phi,t_o,T)$ set it equal to the zero vector in $L_2(R^m)$ and obtain the fact that the optimal control and optimal trajectory for the problem (9) - (10) uniquely satisfy the relation

$$u^m(t,t_o,T,\phi) = u^m(t) = -U^{-1}(t)D^*(t)\int_{t}^{T} X^*(\alpha,t)W(\alpha)x^m(\alpha,t_o,T,\phi)d\alpha \ . \tag{16}$$

Based on (16) we now define the family of mappings, $L(T,t)$, $t_o \leq t \leq T$, by the equation

$$L(T,t)\phi = \int_{t}^{T} X^*(\alpha,t)W(\alpha)x^m(\alpha,t_o,T,\phi)d\alpha \ . \tag{17}$$

This family satisfies the relation $L(T,t)x_t^m(\phi,t_o,T) = L(T,t_o)\phi$. Furthermore, if we let $y(t,\phi) = L(T,t)x_t^m(\phi,t_o,T)$ and $y_t(\phi) = \{y(s,\phi): t \leq s \leq t + h_n\}$ then $y(t,\phi)$ is piecewise differentiable and satisfies (11) for $g(t) = -W(t)x^m(t,t_o,T,\phi)$. Using these facts and relation (15) it is not difficult to obtain the identity

$$\langle y_{t_o}(\phi), x_{t_o}^m(\psi,t_o,T,\phi)\rangle = \int_{t_o}^{T} (W(\alpha)x^m(\alpha, t_o,T,\phi), x^m(\alpha,t_o,T,\phi))d\alpha$$

$$+ \int_{t_o}^{T} (U(\alpha)u^m(\alpha,t_o,T,\phi), u^m(\alpha,t_o,T,\phi))d\alpha = \langle y_{t_o}(\psi), x_{t_o}^m(\phi,t_o,T)\rangle \ . \tag{18}$$

Equation (18) shows that the optimal cost, i.e. the minimum of (9), is given by $\langle y_{t_o}(\phi), x_{t_o}^m(\phi,t_o,T)\rangle$ and that (18) induces a continuous symmetric bilinear form on the space C. From this we can infer the existence for all pairs (t,T), with $t_o \leq t \leq T$, of a family, $g_T(t)$, of continuous linear mappings from C into its topological dual, BV, such that for all ψ and ϕ in C $<g_T(t)\psi,\phi> = <g_T(t)\phi,\psi> = <y_t(\phi),x_t^m(\psi,t_o,T)>$ (see e.g. [7]). We can now easily obtain the analogs of Theorems 1 and 2 of Section 1 for the functional differential problem. Their statements shall be omitted since they are redundant. As in Section 1 we would like to extend our results to infinite time intervals. For this purpose we make the following hypothesis.

Hypothesis 2. For every $\phi \varepsilon C$, independently of $t_o \geqslant 0$, there exists a positive real number $M(\phi)$ such that $C(u,\phi,t_o,T) \leqslant M(\phi)$ for at least one $u \varepsilon L_2(R^m)$.

Necessary and sufficient conditions under which Hypothesis 2 holds for autonomous retarded systems are given in [6]. However, for general functional problems there are no known sufficient conditions other than the assumption that the homogeneous system associated with (10) is uniformly asymptotically stable or that the control space $R^m = R^n$ and $B(t)$ is nonsingular almost everywhere. Thus this problem warrants further investigation. For some results along this line see [2]. In [7] the following results are proved:

Theorem 5. If Hypothesis 2 holds, then for each $t \varepsilon [0,\infty)$ there exists a mapping $g(t)$ from C into BV such that for each pair ϕ,ψ in C, (i) $\lim_{T\to\infty} g_T(t)\phi = g(t)\phi$, (ii) $<g(t)\phi,\psi> = <g(t)\psi,\phi>$ and (iii) $<g(t)\phi,\phi> = C(u^m(t,t_o,\infty,\phi),\phi,t_o,\infty)$. Moreover the optimal control for $T = \infty$ is given by $u^m(t) = -U^{-1}(t)D^*(t)L(t)x_t^m$ where $L(t)$ is the mapping from $C \to R^n$ defined by

$$L(t)\phi = \lim_{s\to 0^-}[(g(t)\phi)(0)-g(t)\phi(s)] \tag{19}$$

Hence the optimal trajectories give rise to the functional differential equation

$$\frac{d}{dt}[x(t)-\sum_{i=1}^{N} A_i x(t-hi)] = \sum_{i=0}^{N} C_i(t)x(t-h_i) - D(t)U^{-1}(t)D^*(t)L(t)x_t. \tag{20}$$

Theorem 6. If Hypothesis 2 holds and $W(t)$ is positive definite, then all solutions of the functional differential equation (20) whose initial conditions are in C tend to the zero vector in C as t tends to infinity.

Corollary. If the problem (9) - (10) is autonomous in the sense that all matrices involved are constant, if W is positive definite and if Hypothesis 2 holds then the optimal trajectories generate a linear evolutionary process in C which is exponentially stable.

3. A DISCUSSION OF THE PROOFS OF THE MAIN RESULTS

The main results in Sections 1 and 2 are obtained in basically the same manner. For $T < \infty$ the cost functionals in both sections are strictly convex positive definite quadratic forms in real Hilbert spaces. Thus we can easily prove they have unique minimizing elements (see e.g. [7]). We compute their Fréchet derivatives and set them equal to the zero vector. Solving the resulting equations for u we argue that these yield the unique optimal controls. This is a consequence of the fact that a necessary condition for a Fréchet differentiable functional to have an extremum at a point u_o is that its derivative evaluated at u_o be equal to the zero vector in the dual space (see e.g. [4]). The second derivatives of the cost functionals have constant values for all u in $L_2(H_2)$. Hence, the cost functionals have one and only one point at which their derivatives are equal to the zero vector in H_2. This establishes the uniqueness of the solutions of equations (4) and (16) and also the linearity of the operators $K(T,t)$ and $L(T,t)$. The fact that $K(T,t)$ induces a symmetric form on H_1 is an almost immediate consequence of the form of $x_u(t,t_o,x_o)$ (e.g.(2)). For functional differential equations we must work harder. Once it is observed that $L(T,t)\phi$ satisfies equation (11), we can use the "adjoint inner product", (14), to obtain the symmetric bilinear functional on $C \times C$ described by equation (18).

To obtain the limiting results it is necessary to prove that the mappings $K(T,t)$ and $g_T(t)$ have limits as T tends to infinity. This is shown by the following theorem which is proved in [6] and is a generalization of a well known result in Riesz-Nagy [10].

Theorem 7. Let $R(t)$, $t \in [0,\infty)$, be a family of continuous symmetric bilinear forms defined on a real Banach space B such that for each $x \in B$

$$\left.\begin{array}{l} R(t)(x,x) \leq R(\bar{t})(x,x) \quad \text{if} \quad t \leq \bar{t} \\ \text{and} \\ \lim\limits_{t\to\infty} R(t)(x,x) < \infty . \end{array}\right] \qquad (21)$$

To each $R(t)$ let there correspond the unique continuous linear mapping $g(t)\colon B \to B'$, the topological dual of B, which satisfies $\langle g(t)x,y\rangle = \langle g(t)y,x\rangle = R(t)(x,y)$. Then there exists a continuous linear mapping $g\colon B \to B'$ such that

$$\left.\begin{array}{l} \langle gx,y\rangle = \langle gy,x\rangle = \lim\limits_{t\to 0} R(t)(x,y) \\ \text{and} \quad \lim\limits_{t\to 0} |g(t)x-gx| = 0 \end{array}\right] \qquad (22)$$

for all x, y in B.

The proof of Theorem 4 is obtained from Theorem 3 and the following theorem which is a version of Theorem 1 in [5].

Theorem 8. An evolutionary process, $S(t,t_o)$, of the type described in Section 1 satisfies an inequality of the form $|S(t,t_o)| \leqslant M_e{}^{\sigma(t-t_o)}$ where $\sigma < 0$ if and only if for each $x \varepsilon H_1$ there exists a constant $M(x)$ such that for all $t_o \geqslant 0$ $\int_{t_o}^{\infty} (S(t,t_o)x, S(t,t_o)x)dt \leqslant M(x)$.

REFERENCES

1. B. D. O. Anderson and J. B. Moore, Linear Optimal Control, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
2. H. T. Banks and M. Q. Jacobs, An attainable sets approach to optimal control of functional differential equations with function space terminal conditions, J. Diff. Eqs., 13(1973), pp. 127-149.
3. R. Bellman and K. L. Cooke, Differential-Difference Equations, Academic Press, New York, 1963.
4. H. Cartan, Calcul Différentiel, Hermann, Paris, 1967.
5. R. Datko, Uniform asymptotic stability of evolutionary processes in a Banach space, SIAM J. Math. Anal., 3(1972), pp. 428-445.
6. R. Datko, Unconstrained control problems with quadratic cost, SIAM, J. Control, 11(1973), pp. 32-52.
7. R. Datko, Some linear nonautonomous control problems with quadratic cost, to appear.
8. N. Dunford and J. T. Schwartz, Linear Operators, Part I, John Wiley, New York, 1958.
9. J. L. Lions, Optimal Control of Systems Governed by Partial Differential Equations, Springer-Verlag, New York, 1971.
10. F. RIESZ and and B. Sz.-Nagy, Functional Analysis, Fredrick Ungar, New York, 1955.
11. D. L. Russell, Quadratic performance criteria in boundary control of linear hyperbolic systems, SIAM J. Cont., 11(1973), pp 475-509.
12. A. P. Sage, Optimum Systems Control, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1968.

MODELS AND CRITERIA IN DISTRIBUTED STABILITY PROBLEMS

J.H. DAVIS
Department of Mathematics, Queen's University, Kingston, Ontario, Canada

ABSTRACT

This paper considers the interplay between models used for physically occurring feedback stability problems and the type of results which are obtained as criteria for stability and instability. Of particular interest is the consistency of stability conclusions as system models are changed, and the theoretical tractability of alternative model types.

INTRODUCTION

This paper discusses some aspects of the problem of input-output stability for certain classes of distributed systems. In particular, various extensions and re-interpretations of the classical Nyquist criterion are discussed. In view of the central role that frequency domain methods have played in the development of stability criteria for systems with finite dimensional realizations, it is of some interest to see to what extent the classical theory generalizes and what role system models play in the final criteria obtained.

Technical details of the results presented below (which may be found elsewhere) are omitted - emphasis is placed instead on the problem framework and its relationship to the results obtained.

We assume the following background knowledge: under assumptions guaranteeing existence of a local unique solution to the loop equation

$$(I + KG)e = u \quad , \tag{1}$$

input-output stability of the feedback system is equivalent to the existence of a bounded inverse for $I + KG$ relative to a Banach space defined on a half-axis in time. [1,2] this formulation of the problem is more suitable to our methods than the equivalent formulation employing the ideas of causality in a central way, [3], and is of wider applicability than the practice of equating stability to the existence of an integrable closed loop impulse response employed in some work on the Nyquist Criterion.

L_1 CONVOLUTION SYSTEMS

Time-invariant systems modelled by ordinary differential equations

$$[p(D) + k\, q(D)]x(t) = u(t) \quad , \quad t \geq 0 \quad , \tag{2}$$

and convolution equations

$$e(t) + k \int_0^t g(t-s)\, e(s)ds = u(t) \quad , \quad t \geq 0 \quad , \tag{3}$$

have input-output stability properties which are of the same nature, in the sense that the stability criterion in each case may be explained in terms of the notion of a Fredholm operator.

Definition:

A closed linear operator T is called a Fredholm operator if (i) $\alpha(T)$ = dim Nullspace of T and (ii) $\beta(T)$ = codim Range of T, are both finite. The index of the Fredholm operator T is $K = \alpha(T) - \beta(T)$.

The basic result in this case is the following:

Theorem 1:

Suppose that (2) and (3) are open loop stable, and that the associated Nuquist locus does not intersect the critical point $-1/k$. Then the minimal closed extension of $p(D) + k\,q(D)$ (in the case of (2)) and the operator $I + kG^*$ (in the case of (3)) are both Fredholm operators of index $K = -\beta(T)$ = encirclements of the Nyquist locus [relative to each of the spaces $L_p(0,\infty)$.]

Basic results leading to the above are in [4,7], and a control theoretic version with unstable open loop is in [2].

The above result provides an interpretation of "why" the system in unstable if the encirclement criterion fails to be satisfied: $I + KG$ has a range space smaller than the class of admissible inputs. In fact, the range reaches a space of finite codimension in the space of admissable inputs. The system is unstable, then, because it is unable to generate an admissable (i.e. bounded in norm) response for an input which lies outside this range space of finite codimension.

In order to connect this interpretation with what is perhaps the more conventional one we make the following remark: the "right half plane poles" of the unstable closed loop make their appearance as the exponents of a decaying exponential function in the nullspace of the adjoint of $I + kG$. These in fact provide the linear functionals which determine the "orthogonality conditions" defining the range of $I + kG$.

SYSTEMS WITH DELAYS

In view of the simple interpretation that the above results may be given, it is natural to ask for which classes of feedback systems such results may be obtained, and what "qualitative features" determine systems for which similar results do hold.

The class of systems consisting of convolution systems with pure delay terms (i.e. almost periodic transfer function) provides an example for which the results are more complicated than the above. The Wiener-Hopf theory of such operators was first discussed in [5] in the case of L_1 kernels, and some related results on open-loop unstable systems are given in [6].

The central aspect of these results is that the stability condition obtained depends on the calculation of an "index" consisting of a real number and an integer.

Theorem 2:

The feedback system described by

$$(4) \qquad e(t) + k \sum_{i=1}^{\infty} a_i \, e(t-\tau_i) + k \int_0^t g(t-s)e(s)ds = u(t) \qquad t \geq 0$$

is input-output stable on $L_2(0,\infty)$ if and only if

(i) the Nyquist locus is uniformly bounded away from the critical point $-1/k$

(ii) the "average number of encirclements" of the Nyquist locus is zero, and

(iii) the number of encirclements of the origin by the locus of

$$1 + \frac{k}{1+k\hat{a}(i\omega)} \cdot \hat{g}(i\omega)$$

is equal to zero.

It is evident from the above that the "delay terms" dominate the stability of the above system in the sense that no additional L_1 kernel will serve stabilize a system which is unstable by virtue of condition (ii) above. In short, the "high frequency" behavior of the transfer function is a dominant factor that no amount of "low frequency" modification can alter.

In the case of an instability of system (4) with only the delay terms present, there is an interpretation of the cause of instability similar to that given above. Instability again results from the fact that the range of $I + KG$ is smaller than the class of admissable inputs, but in this case the range of $I + KG$ is of infinite codimension in the class of admissable inputs. The average winding number measures the size of this codimension [5] in a manner analogous to the previous case.

Of course, the model (4) must be considered an idealization on several grounds: it is probably not physically possible to realize a pure delay, nor is it possible to physically measure a transfer function at arbitrarily high frequencies. It is of interest to see, however, that alternative models incorporating the above objections give the same prediction about the stability of the overall system.

If, as an alternative to (4), one considers the model

$$(5) \qquad e(t) + k \sum_{i=1}^{\infty} a_i \int_0^t \phi(t-\tau_i-s)e(s)ds + k \int_0^t g(t-s)e(s)ds = u(t) \quad , \qquad t \geq 0 \quad ,$$

where $\phi(\cdot) \in L_1$ is a (causal) approximate identity, then the relevant stability criterion is that of Theorem 1. By choosing $\phi(\cdot)$ such that $\hat{\phi} \cong 1$ over a range of frequencies substantially larger than the compact interval where $|\hat{g}| > \epsilon$ (The only circumstance in which (4) and (5) are simultaneously plausible as models), it can easily be seen that Theorem 1 makes the same instability prediction as Theorem 2 . This is true in spite of the fact that certain choices of $\phi(\cdot)$ may cause the Nuquist locus associated with (5) to intersect the critical point $-1/k$, while the model (4) is unstable by virtue of a non-zero average encirclement.

In view of the above and the inherent uncertainty in the model, it would seem advisable to declare (4) "practically" stable if and only if (5) is stable for all "plausible" approximate identities $\phi(\cdot)$.

SYSTEMS WITH DISTRIBUTED FEEDBACK

While the systems discussed above in general have an infinite dimensional realization, the feedback signal is itself scalar valued (analogous results hold for finite-dimensional feedback). We consider in this section results analogous to Theorem 1 which may be obtained for a class of systems with infinite dimensional (i.e. distributed) feedback signals.

The basic model involved is that of a state equation

$$\frac{dx}{dt} = Ax + Bu \tag{6}$$

with A the infinitesimal generator of a stable strongly continuous semigroup of operators on a B-space $\mathfrak{X}$, and bounded operator from a Hilbert space H into $D(A)$. We assume a bounded linear output mapping $C: \mathfrak{X} \to H$, and feedback operator K belonging to the Hilbert-Schmidt class, HS . This results in a closed loop system

$$\begin{aligned} \frac{dt}{dt} &= Ax + Be \\ e &= u - KCx \quad . \end{aligned} \tag{7}$$

Rather than the internal model (7) , we consider the more general input-output model

$$e(t) + \int_0^t W(t-s)e(s)ds = u(t) \quad , \quad t \geq 0 \tag{8}$$

in the space $L_2(0,\infty; H)$, under the assumption that $W(\cdot) \quad L^1(HS)$.

That such models would have a Fredholm theory analogous to Theorem 1 might be suspected on the grounds that the presence of a compact operator in the loop provides a sufficient amount of high frequency filtering around the closed loop.

Theorem 3:

Consider the operator $I + W$ defined in (8), and suppose that the Fredholm determinant $\delta(I + \hat{W})(\cdot)$ does not vanish for real values of its argument. Then $I + W$ is a Fredholm operator, of index $K = -\beta(I+W)$ = number of encirclements of the Nyquist locus of $\delta(I+\hat{W})$. Further (8) is input-output stable if and only if $(I+W)$ is a Fredholm operator of zero index.

Theorem 3 is completely analogous to Theorem 1, in that it identifies instability as the result of the range of $I+W$ having finite codimension in the space of admissable inputs whenever encirclements of the locus occur.

Distributed systems with compact operator-valued transfer functions have been considered in connection with positive operator methods in [8]. The results are of a different nature than Theorem 3.

The discrete analog of Theorem 3 has been used to obtain necessary and sufficient stability conditions for systems with a periodic feedback gain, by casting the problem as a distributed difference equation. [2]. Proof of Theorem 3 and some examples are given in [9].

The usefulness of Theorem 3, of course, is limited by the requirement that (8) be seen as a reasonable model for the problem under consideration. It is clear that models can be substantially helped toward the form (8) by not "over idealizing" the model: spatial domains, if large, <ins>are</ins> genuinely compact, filtering effects <ins>are</ins> present in measuring devices, etc. This area appears to be one in which too much idealization renders the problem <ins>less</ins> tractable, rather than more so.

CONCLUSION

Extensions of the classical Nyquist criterion to classes of distributed systems have been discussed, with emphasis on the role played by the model employed. It has been shown that results which are a logical generalization of the classical case may be obtained, at least in the case of models having a sufficient amount of high frequency antennuation.

REFERENCES

1) J.H. Davis, "Stability Conditions Derived From Spectral Theory; Discrete Systems with Periodic Feedback".

2) J.H. Davis, "Fredholm operators, encirclements, and stability criteria", SIAM Journal on Control, Vol. 10, No. 4, pp. 608-622.

3) J.C. Willems "Stability Instability, Invertibility, and Causality," SIAM Journal on Control, Vol. 7, No. 4, Nov. 1969.

4) M.G. Krein, "Integral Equations on a half-line with kernel depending on the difference of the arguments," AMS Translations, Series II, Vol. 22, pp. 163-288.

5) I.C. Gohberg and I.A. Feldman, "On Wiener Hopf integral - difference equations," Soviet Mathematics Doklady, Vol. 9, (1968) No. 6, pp. 1312-1316.

6) J.H. Davis, "Encirclement conditions for stability and instability of feedback systems with delays," International Journal on Control, Vol. 15, No. 4, pp. 793-799.

7) S. Goldberg, Unbounded Operators, McGraw Hill, New York, 1966.

8) M.I. Freedman, P.L, Falb, and G. Zames, "A Hilbert space stability theory over locally compact abelian groups," SIAM Journal on Control, Vol. 7, No. 3, pp. 479-495.

9) J.H. Davis,"Nyquist conditions for the stability and instability of distributed systems with distributed control", Queen's University preprint.

ON A CLASS OF MULTI-INPUT/MULTI-OUTPUT CONTROL SYSTEMS IN A HILBERT SPACE

PAUL A. FUHRMANN
Division of Engineering and Applied Physics
Harvard University
Cambridge, Massachusetts 02138

ABSTRACT

Recent operator theoretic results, dealing mostly with non self adjoint operators are used to study problems in system theory. The basic idea is to use the universality of the left shift operators in vectorial H^2 spaces in the set of all asymptotically stable contractions in Hilbert space. This idea is used to obtain simple realizations for given multi input/multi output time invariant systems. We study questions of spectral minimality and characterize a class of transfer functions for which the shift realization is spectrally minimal in terms of existence of meromorphic pseudo-continuations.

INTRODUCTION

We consider the input/output relation of a discrete time invariant linear system to be given by a transfer function A which is a $B(\mathbb{C}^p, \mathbb{C}^r)$ valued analytic function in the unit disc D, having a Taylor series expansion $A(z) = \Sigma A_k z^k$. Let X be a Hilbert space, $T \in B(X,X)$, $G \in B(\mathbb{C}^p, X)$ and $H \in B(X, \mathbb{C}^r)$. We say that the system $\{T,G,H\}$ with internal dynamics given by

$$x_{n+1} = Tx_n + Gu_n, \quad y_n = Hx_n \tag{1}$$

is a realization of A if $A_k = HT^kG$ for all $k \geqslant 0$. Thus $A(z) = H(I-zA)^{-1}G$ for all z in D. The discrete system given by (1) is called controllable (observable) if $\bigcap \ker G^*T^{*i} = \{0\}$ $(\bigcap \ker HT^i = \{0\})$. We say that the system $\{T,G,H\}$ is exactly controllable (exactly observable) if the controllability operator $\mathscr{C}$ (the observability operator $\mathscr{O}$) defined on the subset of $\ell^2(0,\infty;\mathbb{C}^p)$ $(\ell^2(0,\infty;\mathbb{C}^r))$ of all finitely nonzero sequences by $\mathscr{C}(\{\xi_k\}) = \Sigma T^k G \xi_k$ $(\mathscr{O}(\{\eta_k\}) = \Sigma T^{*k} H^* \eta_k)$ has a bounded extension to all of $\ell^2(0,\infty;\mathbb{C}^p)$ $(\ell^2(0,\infty;\mathbb{C}^r))$. These notions introduced by Helton are of importance because of their relation to some infinite dimensional ressions of the state space isomorphism theorem [8].

Obviously A(z) is analytic at all points z for which z^{-1} belongs to $\rho(A)$, the resolvent set of A. Let us denote by $\sigma(A) = \{z^{-1} | A \text{ not analytic at } z\}$ then clearly the spectral inclusion property $\sigma(T) \subset \sigma(A)$ holds. We will say that a realization $\{T,G,H\}$ is spectrally minimal if $\sigma(T) = \sigma(A)$ [1]. Since in the context of infinite dimensional systems there is not available a straightforward generalization of a state space isomorphism theorem without introducing extra assumptions on the systems involved then the notion of spectral minimality may be used as a criterion for the usefulness of a given realization.

THE SHIFT REALIZATION

Before proceeding with the solution to the realization problem we introduce some notation related to spaces of vector valued functions. An excellent source is Helson's book [7].

This work was supported by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0006.

Let N,M be two separable Hilbert spaces, to be identified in the sequel with $\mathbb{C}^p$ and $\mathbb{C}^r$. By $\ell^2(0,\infty;N)$ $(\ell^2(-\infty,\infty;N))$ we denote the space of all one sided (two sided) N-valued sequences which have square summable norms. The inner product is defined by $(x,y) = \Sigma(\xi_n,\eta_n)$ for all $x=\{\xi_n\}$, $y=\{\eta_n\}$. Let $\mathbb{T}$ denote the unit circle and $L^2(\mathbb{T};N)$ the space of all N-valued measurable square norm integrable functions on $\mathbb{T}$. The Fourier transform is the unitary map of $\ell^2(-\infty,\infty;N)$ onto $L^2(\mathbb{T};N)$ given by $\mathscr{F}(\{\alpha_n\})=\Sigma\alpha_n e^{int}=a(e^{it})$. Let $H^2(N) = \mathscr{F}\ell^2(0,\infty;N)$, thus $H^2(N)$ is the subspace of $L^2(\mathbb{T};N)$ of all functions whose negative Fourier coefficients vanish. H^2 functions have analytic extensions into the open unit disc from which they can be recaptured as a.e. radial limits. This analytic structure will be freely used and we will denote by the same letters both the boundary value function and its analytic extension. $H^\infty(N,M)$ will denote the Banach space of all $B(N,M)$ valued analytic functions A in D with norm $\|A\|_\infty = \sup\{\|A(z)\| \mid z\varepsilon D\} \leq \infty$. A natural map from $H^\infty(N,M)$ onto $H^\infty(M,N)$ is given by $A \to \tilde{A}$ with $\tilde{A}(z) = A(\bar{z})^*$. The right shift operator S in $H^2(N)$ is defined by $(Sf)(z) = zf(z)$ and its adjoint S^* is given by $(S^*f)(z) = (f(z)-f(0))/z$ and is called the left shift. Subspaces of $H^2(N)$ are called right or left invariant if they are invariant under the right or left shift respectively. Orthogonal complements of right invariant subspaces are left invariant and v.v. The operator J in $L^2(\mathbb{T};N)$ is defined by $(Jf)(e^{it}) = f(e^{-it})$ and is a unitary operator. Given $A \ \varepsilon\ H^\infty(\mathbb{C}^m,\mathbb{C}^n)$ we define the Hankel operator H_A induced by A by $H_A f = P_{H^2(\mathbb{C}^n)}A(Jf)$, where $P_{H^2(\mathbb{C}^n)}$ is the orthogonal projection of $L^2(\mathbb{T};\mathbb{C}^n)$ onto $H^2(\mathbb{C}^n)$. Let $K = \overline{\text{Range } H_A}$ then K is a left invariant subspace of $H^2(\mathbb{C}^n)$. In K we define the restricted right shift T by

$$Tf = P_K(Sf) \qquad \text{for all } f \text{ in } K \tag{2}$$

Here P_K denotes the orthogonal projection of $H^2(\mathbb{C}^n)$ onto K. Clearly $T^* = S^*|K$.

Let now $A \ \varepsilon\ H^\infty(\mathbb{C}^m,\mathbb{C}^n)$, $K = \overline{\text{Range } H_A}$. Define $G : \mathbb{C}^m \to K$ and $H:K \to \mathbb{C}^n$ by $G\xi = A\xi$ and $Hf = f(0)$ respectively, then it is easy to see that $\{T^*,G,H\}$ is a controllable and exactly observable realization of A. This construction is due to Helton [8] and the author in the scalar case [5].

INVARIANT SUBSPACES AND RESTRICTED SHIFTS

To get some more explicit information about the shift realization we have to study the restricted shifts in more detail. The structure of restricted shift operators is determined by the structure of the left invariant subspace. It turns out that the invariant subspaces can be associated with certain bounded operator valued analytic functions in D. We say that a $H^\infty(N,N)$ function Q is rigid if a.e. on $\mathbb{T}$ $Q(e^{it})$ are partial isometries with a fixed initial space. If a.e. $Q(e^{it})$ is unitary then we say Q is inner. Given a rigid function $Q \ \varepsilon\ H^\infty(N,N)$, $QH^2(N)$ is an invariant subspace and all invariant subspaces of $H^2(N)$ have this form. This is the Beurling-Lax theorem [10,6]. Thus the rigid functions give a parameterization of all invariant subspace.

Since the geometry of a left invariant subspace is determined by the rigid function associated with its orthogonal complement, all the spectral information about the restricted shifts is contained in the associated rigid function. The study of these relations started with Moeller [12], Lax and Phillips [11], Helson [7] and Sz.-Nagy and Foias [14]. Thus with the previous notation if K is a left invariant subspace of $H^2(\mathbb{C}^n)$, $K^\perp = QH^2(\mathbb{C}^n)$ and T is defined by (2) then the point spectrum of T^* consists of all points λ, $|\lambda| < 1$ for which $\ker Q(\lambda)^* \neq \{0\}$. If Q is inner the point spectrum of T is the set of all points λ, $|\lambda| < 1$ for which $\ker Q(\lambda) \neq \{0\}$ and the spectrum is the union of the point spectrum and the set of all λ, $|\lambda| = 1$ where $Q(z)$ has no analytic continuation [7].

In case we deal with invariant subspaces associated with inner functions the restricted right and left shifts are closely related. Let Q be inner and let $K = \{QH^2(\mathbb{C}^n)\}^\perp$ and $K = \{\tilde{Q}H^2(\mathbb{C}^n)\}^\perp$ and let T and $\tilde{T}$ be the restricted right shifts in K and $\tilde{K}$ respectively. Define the unitary map τ_Q of $L^2(\mathbb{T};\mathbb{C}^n)$ onto itself by $(\tau f)(e^{it}) = e^{-it}\tilde{Q}(e^{it})f(e^{-it})$ then τ_Q maps K onto $\tilde{K}$ and we have $\tau_Q T^* = \tilde{T}\tau_Q$ [3]. Let $Q_1 \varepsilon H^\infty(\mathbb{C}^n,\mathbb{C}^n)$ and $Q_2 \varepsilon H^\infty(\mathbb{C}^m,\mathbb{C}^n)$ be inner functions and let $K_1 = \{Q_1H^2(\mathbb{C}^n)\}^\perp$, $K_2 = \{Q_2H^2(\mathbb{C}^m)\}^\perp$ and let T_1, T_2 be the restricted right shifts in K_1 and K_2 respectively. A bounded operator $X : K_1 \to K_2$ is said to intertwine T_1 and T_2 if $XT_1 = T_2X$. By a theorem of Sz.-Nagy and Foias [14], refining a result of Sarason [13] every intertwining operator is given by $Xf = P_{K_2}\Gamma f$ for all $f \varepsilon K_1$ with $\Gamma \varepsilon H^\infty(\mathbb{C}^n,\mathbb{C}^m)$ for which $\Gamma Q_1 = Q_2\Omega$ is satisfied for some $\Omega \varepsilon H^\infty(\mathbb{C}^n,\mathbb{C}^m)$ [4].

The spectral properties of intertwining operators for restricted shifts have been studied in [4]. The main tool is the matrix version of the Carleson corona theorem. Given $A_i \varepsilon H^\infty(\mathbb{C}^{n_i},\mathbb{C}^n)$ $(A_i \varepsilon H^\infty(\mathbb{C}^n,\mathbb{C}^{n_i}))$ $i=1,\ldots,p$, the existence of $B_i \varepsilon H^\infty(\mathbb{C}^n,\mathbb{C}^{n_i})$ $(B_i \varepsilon H^\infty(\mathbb{C}^{n_i},\mathbb{C}^n))$ for which $\Sigma B_i(z)A_i(z) = I$ $(\Sigma A_i(z)B_i(z) = I)$ is equivalent to the existence of a $\delta > 0$ such that for all z in D $\inf\{\Sigma||A_i(z)\xi|| \mid \xi \varepsilon \mathbb{C}^n, ||\xi|| = 1\} \geqslant \delta$ $(\inf\{\Sigma||A_i(z)^*\xi|| \mid \xi \varepsilon \mathbb{C}^n, ||\xi|| = 1\} \geqslant \delta)$. This implies that the existence of a bounded left (right) inverse for an intertwining operator X of two restricted shifts is equivalent to the existence of a $\delta > 0$ such that for all z in D $\inf\{||A(z)^*\xi||+||Q_1(z)^*\xi|| \mid \xi \varepsilon \mathbb{C}^m, ||\xi|| = 1\} \geqslant \delta$ $(\inf\{||A(z)\eta||+||Q(z)\eta|| \mid \eta \varepsilon \mathbb{C}^n, ||\eta|| = 1\} \geqslant \delta)$.

CONTROLLABILITY AND OBSERVABILITY

We consider the system $\{T,G,H\}$ with T defined by (2) in a left invariant subspace $K = \{QH^2(\mathbb{C}^n)\}^\perp$ as a state space and $G : \mathbb{C}^p \to K$ and $H : K \to \mathbb{C}^r$ bounded operators. Thus there exist nxp and nxr matrix valued analytic functions for which $(G\xi)(z) = G(z)\xi$ and $(H^*\eta)(z) = H(z)\eta$ for all $\xi \varepsilon \mathbb{C}^p$ and $\eta \varepsilon \mathbb{C}^r$.

Theorem A: a. The system $\{T,G,H\}$ is controllable if and only if G and Q have no common nontrivial left inner factor.

b. The system $\{T,G,H\}$ is exactly controllable if and only if there exists a $\Gamma \varepsilon H^\infty(\mathbb{C}^p,\mathbb{C}^n)$ for which

$$G\xi = P_K\Gamma\xi \qquad \text{for all } \xi \varepsilon \mathbb{C}^p \tag{3}$$

and for which there exists a $\delta > 0$ such that for all z in D

$$\inf\{||\Gamma(z)^*\xi||+||Q(z)^*\xi|| \mid \xi \varepsilon \mathbb{C}^n, ||\xi|| = 1\} \geqslant \delta \tag{4}$$

If we assume $G(z) \varepsilon H^\infty(\mathbb{C}^p,\mathbb{C}^n)$ we can replace Γ by G in (4).

The proof is based on the observation that the Fourier transform of the kernel of the controllability operator is a right invariant subspace of $H^2(\mathbb{C}^r)$. Thus the restriction of the controllability operator to the orthogonal complement is an intertwining operator for two restricted right shifts.

To study the observability of the system we use duality to reduce it to a question of controllability. Thus $\{T,G,H\}$ is (exactly) observable if and only if $\{T^*,H^*,G^*\}$ is (exactly) controllable which is equivalent to the (exact) controllability of the system $\{\tilde{T},\tilde{H},\tilde{G}\}$ in $\tilde{K} = \{\tilde{Q}H^2(\mathbb{C}^n)\}$ with $\tilde{T} = \tau_Q T^* \tau_Q^*$, $\tilde{H} = \tau_Q H^*$ and $\tilde{G} = G^* \tau_Q^*$. Using theorem A we get the following.

Theorem B: a. The system $\{T,G,H\}$ is observable if and only if H_1 and $\tilde{Q}$ have no common nontrivial left inner factor with $H_1(e^{it}) = e^{-it}\tilde{Q}(e^{it})H(e^{-it})$.

b. The system $\{T,G,H\}$ is exactly observable if and only if there exists a $\Lambda \varepsilon H^\infty(\mathbb{C}^r,\mathbb{C}^n)$ satisfying $\Lambda \ker\mathcal{O} \subset QH^2(\mathbb{C}^n)$ for which

$$H^*\eta = P_K\tau^*\Lambda\eta \quad \text{for all } \eta \varepsilon \mathbb{C}^r \tag{5}$$

and for which there exists a $\delta > 0$ such that for all z in D

$$\inf \{||\Lambda(z)^*\eta||+||\tilde{Q}(z)^*\eta|| \mid \eta \varepsilon \mathbb{C}^n, ||\eta|| = 1\} \geqslant \delta \tag{6}$$

In particular the observability operator of the system $\{T,G,H\}$ is the Hankel operator H_A induced by $A(e^{it}) = e^{-it}\tilde{Q}(e^{it})\Lambda(e^{-it})$. Thus the above theorem also answers a question about range closure of Hankel operators.

Next we characterize the class of transfer functions having a realization by exactly controllable and exactly observable systems.

Theorem C: A function $A \varepsilon H^\infty(\mathbb{C}^p,\mathbb{C}^r)$ has a realization by an exactly controllable and exactly observable linear system if and only if it has factorizations of the form

$$A(e^{it}) = e^{-it}Q_{ob}(e^{it})\Omega(e^{it})^* = e^{-it}\Omega_1(e^{it})^*Q_{co}(e^{it})$$

for inner functions $Q_{ob} \varepsilon H^\infty(\mathbb{C}^r,\mathbb{C}^r)$, $Q_{co} \varepsilon H^\infty(\mathbb{C}^p,\mathbb{C}^p)$ and Ω, $\Omega_1 \varepsilon H^\infty(\mathbb{C}^r,\mathbb{C}^p)$ for which there exists a $\delta > 0$ such that for all z in D

$$\inf \{||Q_{ob}(z)\eta||+||\Omega(z)\eta|| \mid \eta \varepsilon \mathbb{C}^r, ||\eta|| = 1\} \geqslant \delta$$

and

$$\inf \{||Q_{co}(z)^*\xi||+||\Omega_1(z)^*\xi|| \mid \xi \varepsilon \mathbb{C}^p, ||\xi|| = 1\} \geqslant \delta$$

FACTORIZATIONS OF OPERATOR VALUED ANALYTIC FUNCTIONS

In the previous section we restricted ourselves to shift systems in left invariant subspaces of the form $\{QH^2(\mathbb{C}^n)\}^\perp$ with inner functions Q. In this section we characterize those functions in $H^\infty(\mathbb{C}^m,\mathbb{C}^n)$ for which the shift realization gives rise to a state space of the above form. Since the state space in the shift realization of a function A in $H^\infty(\mathbb{C}^m,\mathbb{C}^n)$ is the closure of the range of the Hankel operator induced by A we are looking for such A that $\{\text{Range } H_A\}^\perp = QH^2(\mathbb{C}^n)$ for some inner function Q. Such functions will be called strictly noncyclic. The inner function Q associated with A in this way is unique up to a constant unitary factor on the right.

A matrix valued function A is meromorphic of bounded type in a domain V if $A(z) = G(z)/g(z)$ where G is a bounded matrix valued analytic function in V and g a bounded scalar valued analytic function in V. Given $A \varepsilon H^\infty(\mathbb{C}^m,\mathbb{C}^n)$ a $B(\mathbb{C}^m,\mathbb{C}^n)$ valued meromorphic function of bounded type $\hat{A}$ in $D_e = \{z \mid 1 < |z| \leqslant \infty\}$ is said to be a pseudo continuation of A if a.e. on the unit circle $\mathbb{T}$ we have $A(e^{it}) = \hat{A}(e^{it})$. The existence of boundary values follows from Fatou's lemma.

Theorem D: Let $A \varepsilon H^\infty(\mathbb{C}^m,\mathbb{C}^n)$. Then the following statements are equivalent.

(i) A is strictly non cyclic

(ii) A has a meromorphic pseudo-continuation of bounded type in D_e.

(iii) On $\mathbb{T}$ A has a factorization of the form $A(e^{it}) = e^{-it}Q(e^{it})C(e^{it})^*$ for an inner function $Q \in H^\infty(\mathbb{C}^n,\mathbb{C}^n)$ and $C \in H^\infty(\mathbb{C}^n,\mathbb{C}^m)$ such that C, Q have no common nontrivial right inner factor.

(iv) On $\mathbb{T}$ A has a factorization of the form $A(e^{it}) = e^{-it}C_1(e^{it})^*Q_1(e^{it})$ for an inner function $Q_1 \in H^\infty(\mathbb{C}^m,\mathbb{C}^m)$ and $C_1 \in H^\infty(\mathbb{C}^n,\mathbb{C}^m)$ such that C_1, Q_1 have no common nontrivial left inner factor.

The results of this section and the next are generalizations of the work of Kriete [9] and Douglas, Shapiro, and Shields [2] on noncyclic vectors in H^2.

SPECTRAL MINIMALITY

For transfer functions $A \in H^\infty(\mathbb{C}^m,\mathbb{C}^n)$ which are not strictly non cyclic the spectrum of the shift realization generator is the closed unit disc. Thus if $\mathbb{T}$ is not the natural boundary of A the shift realization is not spectrally minimal.

The situation is different when we consider strictly non cyclic transfer functions. The spectrum of the generator, that is the restricted left shift, is determined via the Moeller-Helson theorem from the associated inner function Q. Let $\hat{Q}$ be the meromorphic pseudo-continuation of Q, given by $\hat{Q}(z) = Q(\bar{z}^{-1})^{*-1}$. Thus the singularities of $\hat{Q}$ are at the conjugate points of the zeroes of Q in D and the set of points of $\mathbb{T}$ where no analytic continuation of Q and $\hat{Q}$ exist. Thus we want to study the relation between the singularities of $\hat{Q}$ and those of the meromorphic pseudo-continuation $\hat{A}$ of A. It turns out that $\hat{Q}$ and $\hat{A}$ not only have poles at the same points λ, $|\lambda| > 1$, and of the same orders but also $I_L(A;\lambda) = I_L(Q;\lambda)$ where generally for $|\lambda| > 1$ and a strictly non cyclic A we define $I_L(A;\lambda) = \{P | P \in H^\infty(\mathbb{C}^n,\mathbb{C}^n), P^*A \text{ is analytic at } \lambda\}$. $I_L(A;\lambda)$ is a w*-closed right ideal in $H^\infty(\mathbb{C}^n,\mathbb{C}^n)$. The topology is the one induced by the relation $\{L^1(B(\mathbb{C}^n,\mathbb{C}^m))\}^* = L^\infty(B(\mathbb{C}^m,\mathbb{C}^n))$ via the pairing $(F,G) = \int \operatorname{tr}F(e^{it})G(e^{it})dt$. These ideals are convenient ways of measuring the singularity of $\hat{A}$ at λ. Moreover as far as the boundary behaviour is concerned $\hat{A}$ and $\hat{Q}$ have analytic continuations at the same points of the unit circle. In conclusion we have.

Theorem E: For strictly non cyclic transfer functions in $H^\infty(\mathbb{C}^m,\mathbb{C}^n)$ the shift realization is spectrally minimal.

REFERENCES

1. J.S. Baras and R.W. Brockett, "H^2-functions and Infinite Dimensional Realization Theory," to appear.

2. R.G. Douglas, H.S. Shapiro and A.L. Shields, "Cyclic Vectors and Invariant Subspaces for the Backward Shift Operator," Ann. Inst. Fourier, Grenoble 20 (1970), 37-76.

3. P.A. Fuhrmann, "On the Corona Theorem and its Applications to Spectral Problems in Hilbert Space," Trans. Amer. Math. Soc., 132, (1968), 55-66.

4. P.A. Fuhrmann, "A Functional Calculus in Hilbert Space Based on Operator Valued Analytic Functions," Isr. J. Math., 6 (1968), 267-278.

5. P.A. Fuhrmann, "On Realization of Linear Systems and Applications to Some Questions of Stability," to appear.

6. P.R. Halmos, "Shifts on Hilbert Spaces," J. Reine Augew, Math., 208 (1961), 102-112.

7. H. Helson, Lectures on Invariant Subspaces, Academic Press, N.Y., 1964.

8. J.W. Helton, "Discrete Time Systems, Operator Models and Scattering Theory," to appear.

9. T.L. Kriete, "A Generalized Paley-Wiener Theorem," J. Math. Anal. Appl., 36 (1971), 529-555.

10. P.D. Lax, "Translation Invariant Subspaces," Acta. Math., 101 (1959), 163-178.

11. P.D. Lax and R.S. Phillips, Scattering Theory, Academic Press, N.Y., 1967.

12. J.W. Moeller, "On the Spectra of Some Translation Invariant Subspaces," J. Math. Anal. Appl., 4 (1962), 276-296.

13. D. Sarason, "Generalized Interpolation in H^∞," Trans. Amer. Math. Soc., 127 (1967) 179-203.

14. B. Sz-Nagy and C. Foias, Harmonic Analysis of Operators on Hilbert Space, North Holland, Amsterdam, 1970.

CANONICAL FORMS IN LINEAR SYSTEMS*

BERRIEN MOORE, III
Department of Mathematics, University of New Hampshire, Durham, NH

INTRODUCTION

Given a fixed external description of a linear system, the problem of realization is to give all possible internal descriptions with this external characterization, and to find the relation between them. Of course, one hopes that under natural hypotheses that all of the internal descriptions would be related to some "canonical" realization which would be, in a sense, the simplest possible. It is the purpose of this paper to give a relationship and a canonical realization for a class of shift realized systems. This paper has been greatly influenced by the work of Fuhrmann [4-7] and Helton [11,12]; moreover, the author is particularly grateful for the many stimulating conversations with Professor Fuhrmann.

To set the stage, let $\mathcal{U}$ and $\mathcal{V}$ be complex, separable (either finite or countably infinite dimensional) Hilbert spaces and $\{A_j\}_0^\infty$ a sequence of bounded linear transformations (operators) from $\mathcal{U}$ to $\mathcal{V}$, in notation $A_j \varepsilon \mathcal{B}(\mathcal{U},\mathcal{V})$, then the discrete time invariant input-output map which sends sequences in $\mathcal{U}$ into sequences in $\mathcal{V}$ according to the rule

$$v_n = \sum_{j=0}^{n-1} A_j u_{n-j-1}$$

gives what we have referred to as the external description. The space $\mathcal{U}$ is called the control space, $\mathcal{V}$ the output space, and $\{A_j\}_0^\infty$ the impulse response function.

Now the internal description of an arbitrary discrete constant linear system is given by the dynamical equations

$$x_{n+1} = Tx_n + Bu_n$$

$$v_n = Cx_n$$

where x_n is a vector in another separable, complex Hilbert space $\mathcal{X}$ which is called the state space and $T \varepsilon \mathcal{B}(\mathcal{X},\mathcal{X})$, $B \varepsilon \mathcal{B}(\mathcal{U},\mathcal{X})$, and

---

*Supported in part by NSF Grant GP 14784

$C \varepsilon \mathcal{B}(\mathcal{X},\mathcal{V})$. We shall say that the system $\{T,B,C\}$ realizes $\{A_j\}_0^\infty$ if $A_j = CT^jB$ for all $j \geq 0$. As is mentioned in [5] the realization should satisfy certain minimality requirements. Namely, that it should be both controllable and observable, i.e. that

$$(c) \quad \bigcap_{j=0}^{\infty} \mathrm{Kernel}(B^*T^{*j}) = \{0\}$$

$$(o) \quad \bigcap_{j=0}^{\infty} \mathrm{Kernel}(CT^j) = \{0\} \ ,$$

respectively.

For simplicity let $\mathcal{U} = C^p$ and $\mathcal{V} = C^r$ the p and r dimensional complex Hilbert spaces formed by cross products of the complex plane $\mathcal{C}$. The fundamental result in finite dimensional systems theory (dimension of $\mathcal{X}$ is finite) is that two controllable and observable realizations $\{T_1,B_1,C_1\}$ and $\{T_2,B_2,C_2\}$ realize the same impulse response function if and only if they are similar. To be precise, there exists an invertible operator $E \varepsilon \mathcal{B}(\mathcal{X}_1,\mathcal{X}_2)$ which makes the following diagram commutative

$$\begin{tikzcd}
 & \mathcal{U} \arrow[dl, "B_1"'] \arrow[dr, "B_2"] & \\
\mathcal{X}_1 \arrow[rr, "E"] \arrow[d, "T_1"'] & & \mathcal{X}_2 \arrow[d, "T_2"] \\
\mathcal{X}_1 \arrow[rr, "E"] \arrow[dr, "C_1"'] & & \mathcal{X}_2 \arrow[dl, "C_2"] \\
 & \mathcal{V} &
\end{tikzcd}$$

This is usually referred to as the State Isomorphism Theorem (for instance see [13]). This theorem is of particular importance when viewed against another theorem concerning matrices; namely, that two matrices are similar if and only if they have the same Jordan Canonical Form (see [3; Chapter 7] or [8; Chapter III]).

This result has been extended by Helton [12, Theorem 3b.1] to include infinite dimensional systems; however, one does not obtain, in general, a similarity transform unless the additional assumption that both systems are either exactly controllable or exactly observable (this will be defined in the next section). For a particularly nice discussion of this result see [5].

Motivated by an example of Fuhrmann [7; Section 2] we investigated realizations whose controllability and observability properties are not the same. Specifically, let $\{T_1, B_1, C_1\}$ be controllable and exactly observable and $\{T_2, B_2, C_2\}$ be observable and exactly controllable and suppose T_i is a completely nonunitary contraction [19; p. 8] such that $\mathrm{Rank}(I - T_i^* T_i) = N < \infty$ for $i = 1,2$ (such contractions are said to be of class $C_0(N)$ [19; p. 350] and arise in shift realized systems [1,4-7,11,12]). Under these conditions then, in the last section we shall show that if $\{T_1, B_1, C_1\}$ and $\{T_2, B_2, C_2\}$ have the same impulse response function then T_1 and T_2 are quasi-similar [19, p. 70]. We shall conclude with remarks to indicate that the analogy with the finite dimensional canonical realization is complete.

In the next section we review the necessary results from the theory of Hardy space. For a more complete discussion see Helson [10] or Sz. Nagy and Foias [19].

RESTRICTED SHIFTS

Let $\mathcal{M}$ and $\mathcal{N}$ be separable complex Hilbert spaces, m normalized Lebesgue measure on $[-\pi, \pi)$. Then $L^2(\mathcal{N})$ will denote the Hilbert space of all (equivalence classes) weakly measurable functions from the unit circle to $\mathcal{N}$ having square-integrable norm with the inner product in $L^2(\mathcal{N})$ given by

$$< f,g > = \int_{-\pi}^{\pi} < f(e^{it}), g(e^{it}) >_{\mathcal{N}} dm(t) .$$

The Hardy subspace $H^2(\mathcal{N})$ of $L^2(\mathcal{N})$ is the closed subspace of all functions whose negative indexed Fourier coefficients are all zero. We remark that the functions of $H^2(\mathcal{N})$ have analytic extensions into the open unit disk from which they can be recaptured almost everwhere as radial limits. If we let χ denote the identity function on the unit circle then the unilateral shift S on $H^2(\mathcal{N})$ is given by

$$S: \ f \to \chi f .$$

Closely connected with $H^2(\mathcal{N})$ is the sequence space

$$\ell^2(\mathcal{N}) = \{\{x_j\} | \alpha_j \in \mathcal{N} \text{ and } \sum_{j=0}^{\infty} ||\alpha_j|| < \infty\} .$$

The inner product is given in the obvious manner. Further there is a unitary map (Fourier transform) which identifies the $\ell^2(\mathcal{N})$ sequence as the nonnegative Fourier coefficients of some $f \in H^2(\mathcal{N})$.

Let $L^\infty(\mathcal{B}(\mathcal{M},\mathcal{N}))$ be the space of weakly measurable, essentially bounded functions from the unit circle to $\mathcal{B}(\mathcal{M},\mathcal{N})$ equipped with the sup norm. A function $A(\cdot)$ in $L^\infty(\mathcal{B}(\mathcal{M},\mathcal{N}))$ is said to be analytic (note--a slight misuse of the term since $A(\cdot)$ is considered to be defined on the unit circle) if

$$\int_{-\pi}^{\pi} < A(e^{it})m,n >_{\mathcal{N}} e^{ijt} dm(t) = 0$$

whenever $m \varepsilon \mathcal{M}$, $n \varepsilon \mathcal{N}$, and $j = 1,2,3,\cdots$. The space of analytic functions in $L^\infty(\mathcal{B}(\mathcal{M},\mathcal{N}))$ is denoted by $H^\infty(\mathcal{B}(\mathcal{M},\mathcal{N}))$. A function in $H^\infty(\mathcal{B}(\mathcal{M},\mathcal{N}))$ can be realized as a weakly analytic function on the open unit disc where again the norm is given by supremum. We shall blur the distinction between these two spaces and use the same notation for either.

Given any $A(\cdot) \varepsilon L^\infty(\mathcal{B}(\mathcal{M},\mathcal{N}))$, then we can define the Hankel operator associated with A on $H^2(\mathcal{M})$ to $H^2(\mathcal{N})$ by

$$H_A f = P(ARf)$$

where $R: L^2(\mathcal{M}) \to L^2(\mathcal{M})$ is defined by $(Rf)(e^{it}) = f(e^{-it})$ and P is the orthogonal projection of $L^2(\mathcal{N})$ onto $H^2(\mathcal{N})$. It is not difficult to check that

$$S^* H_A = H_A S .$$

The notation is somewhat imprecise since S denotes the unilateral shift in $H^2(\mathcal{M})$ and S^* the adjoint of the unilateral shift in $H^2(\mathcal{N})$. Consequently, the range closure of a Hankel operator is invariant under the adjoint of the unilateral shift, or equivalently the orthogonal complement is invariant under S. The closed invariant subspaces of S in H^2 were described by Beurling [2]. Specifically, if $\mathcal{Z} \subseteq H^2$ is S-invariant then there exists $g \varepsilon H^\infty$ such that $|g(e^{it})| = 1$ almost everywhere and $\mathcal{Z} = gH^2$. In higher dimensions the result is similar ([9],[14],[18]; see also [10] or [19]). Namely, if $\mathcal{Z} \subseteq H^2(\mathcal{N})$ is S-invariant then there exists a function $\Theta(\cdot) \varepsilon H^\infty(\mathcal{B}(\mathcal{N},\mathcal{N}))$ such that $\Theta(e^{it})$ is a partial isometry almost everywhere with a fixed initial space and $\mathcal{Z} = H^2(\mathcal{N})$. When $\mathcal{N}$ is finite dimensional and Θ is almost everywhere an isometry (or equivalently unitary, in this case) then Θ is called inner. This is the same as demanding the S-invariant subspace to be of full range [10].

Finally, let $\mathcal{Z}$ be an S-invariant subspace of full range in $H^2(\mathcal{N})$ where the dimension of $\mathcal{N}$ is N, then define the operator $S(\Theta)$ on the orthogonal complement $\mathcal{H}(\Theta)$ of $\Theta H^2(\mathcal{N}) = \mathcal{Z}$ in $H^2(\mathcal{N})$ by

$$S(\Theta)f = P\,(Sf)$$

where $f \in \mathcal{H}(\Theta)$ and P_Θ is the orthogonal projection of $H^2(\mathcal{N})$ onto $\mathcal{H}(\Theta)$. We remark that any $C_0(N)$ operator is unitarily equivalent to some $S(\Theta)$ [19].

With these facts in hand we return to the problem of realization and briefly review what is now known, in particular we rest this discussion on the recent work of Fuhrmann [4-7] and Helton [11,12] about shift realized systems. First, we define exact controllability and exact observability which was mentioned in the introduction. Let Δ be the dense subset of all finite sequences in the space $\ell^2(\mathcal{C}^p)$. We define the controllability operator K_c of the system $\{T,B,C\}$ where $T \in \mathcal{B}(\mathcal{X},\mathcal{X})$, $B \in \mathcal{B}(\mathcal{C}^p,\mathcal{X})$ and $C \in \mathcal{B}(\mathcal{X},\mathcal{C}^r)$ by

$$K_c(\{\alpha_j\}) = \sum_{j=0}^{\infty} T^j B\alpha_j$$

where $\{\alpha_j\} \in \Delta$. We say that the system is exactly controllable if K_c can be extended to an operator of $\ell^2(\mathcal{C}^p)$ onto $\mathcal{X}$.

In a like manner we define the observability operator K_0 on Δ' to X where Δ' is the similar dense set but in $\ell^2(\mathcal{C}^r)$ by

$$K_0(\{\beta_j\}) = \sum_{j=0}^{\infty} T^{*^j} C^* \beta_j$$

for $\{\beta_j\} \in \Delta'$. The system is exactly observable if K_0 can be extended to an operator of $\ell^2(\mathcal{C}^r)$ onto X.

Now, let $\{A_j\}_0^\infty$ be an impulse response function with $A_j \in \mathcal{B}(\mathcal{C}^p,\mathcal{C}^r)$ such that the function $A(\cdot)$ defined on the unit circle by

$$A(e^{it}) = \sum_{j=0}^{\infty} A_j e^{ijt}$$

is in $H^\infty(\mathcal{B}(\mathcal{C}^p,\mathcal{C}^r))$. This function (or more accurately its analytic extension into the open unit disc) is usually called the transfer function. Let H_A be the Hankel operator from $H^2(\mathcal{C}^p)$ to $H^2(\mathcal{C}^r)$ induced by A. Suppose $[\Theta H^2(\mathcal{C}^r)] = \overline{\text{Range}(H_A)} = \mathcal{H}(\Theta)$ is a nontrivial S^*-invariant subspace such that Θ is inner, and consider the operators $B \in \mathcal{B}(\mathcal{C}^p,\mathcal{H}(\Theta))$ and $C \in \mathcal{B}(\mathcal{H}(\Theta),\mathcal{C}^r)$ defined by

$$B\alpha = A\alpha \ , \quad \alpha \in \mathcal{C}^p \quad \text{and}$$
$$Cf = f(0) \ , \quad f \in \mathcal{H}(\Theta) \ .$$

Note that if one consider α as a constant function in $H^2(\mathcal{C}^p)$ then $A\alpha = H_A(\alpha)$. Under these definitions the system $\{S^*(\Theta),B,C\}$ realizes $\{A_j\}_0^\infty$ and is controllable and exactly observable ([12], see also [5-7]);

however, Fuhrmann [7] has given another shift realization $\{S(\Theta_1),B_1,C_1\}$ for the given impulse response function $\{A_j\}_0^\infty$ such that the system is observable and exactly controllable. It is to the question of the relationship between such unmatched realizations that we turn to in our final section.

CANONICAL REALIZATION FOR RESTRICTED SHIFT SYSTEMS

We begin with a definition which was implicit in the concluding remarks of our previous section.

3.1 Definition. A multiple input--multiple output restricted shift system is a triple $\{T,B,C\}$ where T is a $C_0(r)$ contraction operator on a Hilbert space X , $B \in \mathcal{B}(C^p,X)$, and $C \in \mathcal{B}(X,C^r)$.

This definition though phrased in terms of C_0 operator with finite defect is equivalent to Definition 3.1 of [6].

Finally, we introduce the concepts of quasi-equivalence of matrices over H^∞ and quasi-similarity of operators on a Hilbert space X . Quasi-equivalence was introduced by Nordgren [16], but we shall use an equivalent definition which is slightly easier to state given in a later paper by Szücs [20]; see also [17].

3.2 Definition. A matrix Θ_1 over H^∞ is said to be quasi-equivalent to a matrix Θ_2 over H^∞ if and only if for every nonzero $\psi \in H^\infty$ there are two square matrices Γ_1 and Γ_2 such that $\det \Gamma_1$ and $\det \Gamma_2$ are both relatively prime to ψ (no nontrivial common inner divisors) and $\Gamma_1\Theta_1 = \Theta_2\Gamma_2$.

And finally,

3.3 Definition [19, p. 70]. A quasi-affinity from X_1 to X_2 we mean an operator $E \in \mathcal{B}(X_1,X_2)$ which is one-to-one and has dense range. If $T_1 \in \mathcal{B}(X_1)$ and $T_2 \in \mathcal{B}(X_2)$ and there exists a quasi-affinity $E \in \mathcal{B}(X_1,X_2)$ such that $T_2E = ET_1$ the T_1 is said to be a quasi-affine transform of T_2 . The operators T_1 and T_2 are called quasi-similar if they are quasi-affine transforms of one another.

These two concepts were connected in [15] for $C_0(r)$ operators in the following:

3.4 Theorem. Let $S(\Theta_1)$ and $S(\Theta_2)$ be arbitrary $C_0(r)$ operators with Θ_1 and Θ_2 $r \times r$ matrices over H^∞ such that $\Theta_1(e^{it})$ and $\Theta_2(e^{it})$ are unitary valued (i.e., $\Theta_i(\cdot)$ are inner functions in $H^\infty(\mathcal{B}(C^r,C^r))$, $i = 1,2.$). $S(\Theta_1)$ and $S(\Theta_2)$ are quasi-similar if and only if Θ_1 and Θ_2 are quasi-equivalent.

This result is of particular interest when viewed against two results:

(a) Sz.-Nagy and Foias (for a discussion see [19, p. 369-71]) prove that any $S(\Theta) \varepsilon C_0(r)$ operator is quasi-similar to a Jordan operator, i.e., an operator of the form $S(m_1) \oplus \cdots \oplus S(m_r)$ where m_i is an inner function in H^∞ for $i = 1,2,\cdots,r$.

(b) [16] Every matrix Θ over H^∞ is quasi-equivalent to a diagonal matrix where the entries on the diagonal are the invariant factors of Θ . Sz.-Nagy and Foias' works showed that the minimal function [19; Chapter IX] of a $C_0(r)$ contraction $S(\Theta)$ bears the same relation to the inner matrix function Θ (actually to the characteristic operator function but the difference is nonessential) that the minimal polynomial of a finite matrix M over $\mathcal{C}$ bears to the polynomial matrix $M - \lambda$. The work in [15] and [16] showed that the invariant factors of Θ determine the Jordan model of $S(\Theta)$. This rather complete analogy with finite matrices is carried further in the following generalization of the State Space Isomorphism Theorem for unmatched restricted shift systems.

3.5 Theorem. If $\{S(\Theta_1),B_1,C_1\}$ and $\{S(\Theta_2),B_2,C_2\}$ are restricted shift systems such that $\{S(\Theta_1),B_1,C_1\}$ is controllable and exactly observable and $\{S(\Theta_2),B_2,C_2\}$ is observable and exactly controllable where $S(\Theta_i) \varepsilon \mathcal{B}(\mathcal{H}(\Theta_i),\mathcal{H}(\Theta_i))$, $B_i \varepsilon \mathcal{B}(\mathcal{C}^p,\mathcal{H}(\Theta_i))$, and $C_i \varepsilon \mathcal{B}(\mathcal{H}(\Theta_i),\mathcal{C}^r)$ for $i = 1,2$. If $\{S(\Theta_1),B_1,C_1\}$ and $\{S(\Theta_2),B_2,C_2\}$ have the same impulse response function, then $S(\Theta_1)$ and $S(\Theta_2)$ are quasi-similar. Consequently, $S(\Theta_1)$ and $S(\Theta_2)$ have the same Jordan model.

Proof. This proof follows very closely Helton's proof of Theorem 3.1.b in [12]; therefore, we will only sketch the ideas. Since $\{S(\Theta_1),B_1,C_1\}$ and $\{S(\Theta_2),B_2,C_2\}$ have the same impulse response function we have

$$K^*_{1_o} S^n(\Theta_1)K_{1_c} = K^*_{2_o} S^n(\Theta_2)K_{2_c}$$

for $n = 0,1,2,\cdots$. The controllability and observability requirements in our hypotheses implies that

$$K^{*^{-1}}_{1_o} K^*_{2_o} = K_{1_c} K^{-1}_{2_c}$$

and that each is bounded, one-to-one, and has dense range. Consequently,

$$E = K_{1_c} K^{-1}_{2_c}$$

is a quasi-affinity from $\mathcal{H}(\Theta_2)$ to $\mathcal{H}(\Theta_1)$ such that

$$S(\Theta_1)E = ES(\Theta_2)$$

thus $S(\Theta_2)$ is a quasi-affine transform of $S(\Theta_1)$. It is at this point that we use (for the first time) the fact that the systems arise as restricted shifts. By Theorem 4 of [15] we conclude that Θ_1 and Θ_2 are quasi-equivalent. Theorem 3.4 gives the conclusion.

3.6 Remarks. The analogy with finite dimensional systems is even more complete since:

(a) The spectrum $C_0(r)$ operator is invariant under quasi-similarity (for a negative result see [5; Appendix 1]).

(b) The determinant of matrices over H^∞ is invariant under quasi-equivalence.

REFERENCES

1. J.S. Baras and R.W. Brockett, "H^2-functions and Infinite Dimensional Realization Theory," (to appear).
2. A. Beurling, "On Two Problems Concerning Linear Transformations in Hilbert Space," Acta. Math., 81(1949), 239-255.
3. D.T. Finkbeiner, II, Introduction to Matrices and Linear Transformations, W.H.Freeman, San Francisco, 1965.
4. P.A. Fuhrmann, "On Realization of Linear Systems and Applications to Some Questions of Stability," (to appear), Math. Sys. Th.
5. P.A. Fuhrmann, "Exact Controllability and Observability and Realization Theory in Hilbert Space," (to appear).
6. P.A. Fuhrmann, "Realization Theory in Hilbert Space for a Class of Transfer Functions," (to appear).
7. P.A. Fuhrmann, "Factorization Theorems for a Class of Bounded Measurable Operator Valued Functions," (to appear).
8. I.M. Gelfand, Lectures on Linear Algebra, Interscience, New York, 1961.
9. P.R. Halmos, "Shifts on Hilbert Spaces," J. Reime Augew, Math., 208 (1961), 102-112.
10. H. Helson, Lectures on Invariant Subspaces, Academic Press, N.Y., 1964.
11. J.W. Helton, "The Characteristic Functions of Operator Theory and Electrical Network Realizations," Indiana Univ. J. Math., 22(1972), 403-414.
12. J.W. Helton, "Discrete Time Systems, Operator Models and Scattering Theory," (to appear).
13. R.E. Kalman, P.L. Falb, and M.A. Arbib, Topics in Mathematical System Theory, McGraw-Hill, New York, 1969.
14. P.D. Lax, "Translation Invariant Subspaces," Acta Math., 101(1959), 163-178.

15. B. Moore, III and E.A. Nordgren, "On Quasi-equivalence and Quasi-similarity," Acta Sci. Math., 34(1973), 311-316.

16. E.A. Nordgren, "On Quasi-equivalence of Matrices over H^∞," Acta Sci. Math., 34(1973), 302-310.

17. E.A. Nordgren, "The Ring N^+ is Adequate," (to appear), Acta Sci. Math.

18. J. Rovnyak, "Ideals of Square Summable Power Series," Proc. Amer. Math. Soc., 13(1962), 360-365.

19. B.Sz-Nagy and C. Foias, Harmonic Analysis of Operators on Hilbert Space, Akadémiai Kiadó, Budapest, 1970.

20. J. Szücs, "Diagonalization Theorems for Matrices Over Certain Domains," (to appear), Acta Sci. Math.

TRENDS AND PHILOSOPHIES OF POWER SYSTEMS CONTROL

OLLE I. ELGERD
Department of Electrical Engineering
University of Florida
Gainesville, Florida 32611

ABSTRACT

Due to the extreme complexity of even the simplest power systems the philosophy of decomposition has been and will continue to be a fundamental basis for design of the control systems. The availability of the computer is drastically changing the role of the human operator. On-line computation of load flow, power angle and frequency dynamics are predicted to increase system security in spite of geometrically growing system size. The computer will also take over more of the decision making job.

INTRODUCTION

By "control" of power systems we shall understand in this paper all the varied operating functions the objective of which is to:

1. keep the system operating in its "normal" or "preventative" state
2. limit to a minimum the extent of damage to the system during an "emergency" or "fault" state
3. and--finally--restore the system to normalcy during the "restorative" or "post-fault" state.

It is unrealistic to assume that we ever shall be able to develop a central super control computer which will have instantaneous and complete control over all variables which characterize the operating state of a power system. For a systems engineer who has seen how, in a few seconds, a simple line short-circuit can cascade into a complete blackout the thoughts of such a computer may be a comforting but unfortunately very unrealistic pipe dream.

It is of interest to summarize here some of the unique features which make a modern power system such a challenge to a control systems engineer:

1. Size and complexity. Modern power systems represent (measured in invested capital) the largest man-made systems in existence. A control engineer measures "complexity" in terms of dimensionality or number of state variables. A typical power system may require thousands of state variables for adequate description.
2. The changing character of the systems on a daily and yearly scale. The normal operating states compared say between noon and midnight are entirely different. The exponential load growth of a typical system presently corresponds to a doubling time of about 9-10 years. In the next decade we must probably add as much generating capacity as we have today accumulated since the beginning of the technological age. (The energy crisis may force us to substantially reduce the growth rate in decades to follow.)
3. Due to the 24 hour operation it is practically impossible to conduct meaningful full-scale systems tests. For example, we still have been unable to test the validity of composite load models during extensive voltage swings.
4. The geographical spread of a typical system makes it vulnerable to environmental influences of statistical nature, and makes data

transmission difficult.

5. Control inputs are mixed continuous--discrete.
6. System dynamics range over a very broad bandwidth. Line transients may last for milli- and micro-seconds. Short circuit phenomena range from milliseconds to seconds. Generator rotor swings and frequency transients may take seconds to minutes.

Let us consider how the above facts have contributed to shape the present control philosophies. It should be added that just as there exists no two identical power systems there likewise does not exist two identical power systems control schemes. Also, it must be borne in mind that rarely has there been a very drastic change in power system control technology. The changes that do take place are of the same steady evolutionary type as characterizes the growth of the systems themselves. This trend will most probably continue into the future.

PRESENT CONTROL PRACTICES

Control In The Normal State - Pf And QV Control

Typically, an electric power system is being maintained in its "normal" mode or state in excess of 99 percent of the time, and the control systems which strive to maintain this state therefore occupy a central position in our discussion. The normal mode of operation is characterized by the following set of *primary* control requirements:

1. The power demands must be met at all times at the various load points.
2. The system frequency must be maintained within very rigid limits.
3. A specified voltage profile must be kept throughout the network.

If the system is part of a power pool a fourth primary requirement is to keep the tie-line powers (i.e. the power exchange with the neighbors) at certain fixed values.

The two first requirements are met by maintaining a continuous close balance between generated and demanded *real* power. When such a balance is kept, frequency constancy is achieved and the frequency error is in fact the best balance indicator. We refer to this as *load frequency control* (LFC) or *megawatt-frequency* (Pf) control.

The third operating objective is obtained by keeping balance between demanded and generated *reactive* power. If such a balance is not kept voltage fluctuations will result and we can thus utilize the voltage errors as indicators that the *megavar-voltage* (QV) control is obtained.

The fourth operating requirement is met by adding to the control signal of the LFC system an error signal representing the deviations in the contracted power exchange to neighboring pool members.

Frequency and voltage control are thus *primary* control functions in all power systems. As the accuracy requirements are much higher for frequency than for voltage we can probably state that LFC is the most important single power system control function.

This author has suggested a simple LFC analog.(1) A train consisting of a number of engines (representing the generators in the electric system) and freight cars (representing the loads) interconnected by elastic couplings (representing transmission lines) constitutes in effect a very true dynamic model of an electric power system from an LFC point of view. Keeping the train travelling at a constant speed by means of controlling the engine powers is directly analogous to keeping the power system frequency constant by controlling the generator powers.

Pf and QV control are today performed *automatically*, and in *closed loop* fashion. The fact that today's power networks in spite of their complexity are performing as well as records prove to be true is due to the reliability of these primary control functions. There are two basic reasons for this fortunate circumstance:

1. The Pf and QV control channels are essentially *non-interacting*. Manipulation of the control forces in the former (turbine torques) has little or no effect upon the state variables in the latter (voltage magnitudes). Similarly, the manipulation of the control forces in the latter (excitation voltages) has little effect upon frequency.
2. The QV controls of individual geographically dispersed generator units interact only weakly.

Control Through Decomposition

By thus introducing the two fundamental primary control jobs, Pf and QV control, we have in fact encountered the fundamental power systems control philosophy--*control system design by decomposition*. This is, of course, nothing unique for power systems but it is probably true that in no other industry does the decomposition philosophy so thoroughly permeate the control picture.

Decomposition simply means that we accept from the outset the impossibility of designing a single central control system that would handle the overall control job. Instead, we subdivide the overall controls job into "pieces" which are simple enough to lend themselves to reasonably simple analysis and understanding.

The present and future success of power system operation will depend greatly upon the skill with which we shall be able to master the art of slicing the geometrically growing power systems into conceptually digestible pieces.

Decomposition is performed in many dimensions. Pf and QV channels exemplified the utilization of the decoupling features between the real and reactive power controls. Sometimes such natural lines of division can be found intuitively, sometimes by analysis. Sensitivity analysis is particularly useful in this regard.

Geographical and functional decomposition bases have been used for many years. For example, the various controls of a generator unit are physically located in the generator station in question. The voltage level for a certain distribution area is primarily controlled by tap changing of the transformer serving the area in question, etc.

We have already divided the operating states into "normal", "emergency", and "restorative". It is natural to divide the control systems along the same lines. Pf and QV controllers belonged in the first group. Automatic switching, system splitting, and fault clearing, etc., are "emergency" controls, and automatic feeder restoration, load transfer, and load shedding are "restorative" controls.

Time response has long been an important decomposition basis in power system control design, particularly in the "normal" and "emergency" modes of operation. The ultra-fast line surges caused by atmospheric disturbances are controlled within microseconds by means of fast acting lightning arresters. The surges will as a rule not operate the slower line protection relays. However, should the surges destroy the line insulation with resulting short circuits then these relays go into operation. Usually we have several "defense lines" divided on a time base. This time base decomposition is of particular importance in the normal state control. The Pf and QV controllers operate within the seconds region in their attempts to keep the system in the chosen state. This particular state is, however, not unique. It is usually chosen on basis of some optimum criterion (minimum fuel cost for example). The computation of such an optimum operating state is a complex computational process that in a practical situation may not be possible to achieve on a more frequent basis than say every fifteen minutes. We evidently identify here a hierarchy of control levels. The primary (faster) level maintains continuously the chosen operating state which has been computed in a secondary (slower) control loop the purpose of which it is to maintain an optimum operation in spite of the slowly shifting load demands.

ROLE OF HUMAN OPERATOR

Man has always been an integral part of the control of power systems. Long before the Pf control loops went automatic the human operator kept vigil over the system frequency and commanded appropriate position changes to the generator turbine valves. As more and more of the various subcontrols went automatic man's role changed accordingly but he was never fully replaced by machines.

Even the entrance on the scene by the computer did not render man superfluous. And nothing in the present picture or that of the foreseeable future will change this fact. However, the role of the human operator is undergoing drastic changes. As most of the above mentioned control jobs during the normal mode of operation are entirely automatic and as this mode as earlier stated prevails for more than 99 percent of the time it would seem to indicate that the human operator has become a passive observer.

Recently a pilot of the highly computerized Boeing 747 airliner was quoted:

> "Modern airline piloting consists of hours of boredom interrupted by seconds of sheer terror."

Could this be a fair representation of the job description of the operator of a modern power system?

Not so--because the availability and proper use of the computer makes it possible for the human operator to *predict* what the system is going to do in the minutes ahead and take appropriate *anticipatory* action. In fact this constant vigil over the system trends and the formulation of the best strategy decisions that must be made to meet possible contingencies will be the main functions of the human operator--and it must be noted that these matters will continuously occupy his mind during normal operation. It is true that some of the decisions can be left to the computer but no computer can ever match the ability of the human mind to make intelligent choices when a myriad of solution possibilities exist.

The Dispatch Center

Although no general rules govern the methods whereby power systems integrate their control functions they all have many features in common--in spite of their widely different sizes.(2,3) All have a central control facility usually referred to as a "dispatch center" (DC).

The DC typically contains a large display map showing transmission network, generator stations and bulk power substations. All important components are indicated with symbols, giving the operator instantaneous visual information of system interconnection status. Chart recorders for individual generations, tie-line flows, frequency, loads, etc. usually are standard equipment.

Few modern DC's lack a large computer facility. This computer may perform many simultaneous tasks, depending upon the sophistication of the system in question and depending greatly upon the quality of the data transmission system in use.

The data transmission method critically influences the use of the computer and the overall efficiency of the DC in general. The ideal situation--never to be found--would be one in which continuous streams of all system state variables would reach the DC. In reality the data systems range from oral phone communications to high-speed digital microwave links. More will be said about this important matter in next section.

Based upon available data the computer typically would perform the following tasks:

1. As part of the LFC loop it would compute and dispatch appropriate generation commands to the generator units under control.
2. Compute optimum operating points ("Economic dispatch").
3. Monitor (maybe on a 5 min. scan) line loads.
4. Perform load forecasting.
5. Compute unit commitment.
6. Pool scheduling.
7. Reserve analysis.
8. Data logging.

The DC operator (and by "operator" we shall mean here the staff who have direct control of the monitoring and control of the power system) will in general be exercising control of the system in three different operating modes; *preventative, corrective,* and *restorative.*

The *preventative* mode corresponds to the "normal" mode of system operation. It indicates that although the system may operate "normally" the operator must still maintain a watchful eye on the progress. To aid the operator in his decisions he has available all the data outputs from charts, display and the computer programs whatever they may be. In the preventative mode the operator is influenced by the *predicted* system state rather than the actual one. Typical preventative operator functions may include:

a) Modification of unit commitment and economic dispatch parameters to alter the spinning reserve.

b) Modification of the logic functions that would control the system separation in case of emergency.
c) Modification of automatic load shedding schedules.

Should the system enter the "emergency" or "fault" state the operator will work in the *corrective* mode. His actions now might include:

a) Start-up of fast peaking units.
b) Dropping interruptable loads.
c) Rescheduling of generation etc.

In this mode of operation the operator's decisions could be greatly influenced by the computer, which can fast assimilate data and present the operator with information that would aid him in speedy decision making.

In the restorative mode of operation the causes of the emergency have been eliminated or at least isolated and the operator can now concentrate on (again aided by the computer) to get the system quickly back to "normal" operation.

FUTURE TRENDS

The above represents a very sketchy discussion of the present status of power systems control. We have tried to give the reader a brief insight into the control problem.

We will try now to look into the crystal ball and predict how this technology may evolve in the future. We can distinguish several probable developments directions.

Improved Data Acquisition And State Estimation

Without doubt the future DC operator will be able to make his decisions based upon more reliable data than he has available today. Large efforts by many people are today channeled into the area of "state estimation". This is a general description of the theory and techniques of assembling, filtering, estimating, and interpreting data. However sophisticated a control system is its quality of operation will always be determined by the completeness and accuracy of the input data. The data gathering process has always been a weak link in power systems because of distances involved, amount of data, noise corruption and ignorance about what data really is significant.

The latter point is important indeed. There is no reason at all to transmit to the DC every physical variable throughout the system. For example, one of the most useful pieces of information the operator can have is the updated power flow situation throughout the grid. The point we wish to make here is that this flow picture can be *computed* in the DC based upon a *limited* set of variables.

Developments in the very active "state estimation" field are therefore of acute interest. This author feels that there presently maybe is too much emphasis on state estimation per se and not enough work in the area of what to do with this data once it will be available. There certainly is much opportunity here for the potential researcher.

"Decentralization" Of Controls Job

We stated earlier the importance of decomposition. Certainly we shall see further trend in this direction. There is a human characteristic to create bureaucracies--and as we build more and more sophistication into the future DC we may be tempted to charge it with more functions and responsibility than we should. "Local control" will never lose its usefulness.

Generator stations, substations, and local load centers will have their own small mini-computers which will process the local data and perform local controls forwarding only the bare minimum of information to the DC.

On-Line Load-Flow Displays

An up-to-date picture of the system load flow is as important for the DC operator as is the traffic volume data for a traffic controller. No doubt this will be a "must" in all future power systems control centers. The line flow can be presented visually to the operator in many ways the best method probably being as an overlay on the wall display. Present "fast load-flow" computation techniques are available to make this a reality, but the computer must be presented with the proper generation data, real and reactive.

If the operator has a minute-by-minute picture of the load flow he (or the computer) can spot the critical lines where load or thermal limits are being reached and thus be in a better position to take proper preventative actions.

Dynamic Rescheduling And Control

Assume that the operator notices that the line flow approaches a configuration that may result in several "soft spots", i.e., overheated or overloaded lines or a situation when dynamic stability is impaired. There is a need now to reschedule the generators so as to obtain a new and better load flow configuration. How to do this in time? The problem is referred to as dynamic rescheduling. We need presently better algorithms for this job and, particularly, we need to know how to seek "better" load flows and how to define "better". Clearly the rescheduling rates are limited not only by the computer but also by the physical response limits of boiler-turbines.

Dynamic Stability Assessment By Means Of Real Time Computation Of Swing Curves

One of the most important functions of a power systems operator is to try to prevent the system from drifting into an operating state characterized by poor stability margins. Should a fault occur while the system is in such a state the operator would be unable to prevent the system from cascading into a blackout.

The way we presently handle this situation is by precomputing a large number of contingency cases. The computations consist of integrating usually very complicated sets of coupled differential equations in order to obtain so-called "swing curves", i.e. time plots of generator rotor angles following a fault. Upon precomputing a large number of cases we obtain a feel for what type of configurations and faults that would prove dangerous and we then instruct the operator to try to avoid getting close to these contingencies. Clearly this is not a very good procedure. For one thing we can never precompute all possible cases.

A better way to handle this problem would be to let the computer run *in real time* stability checks on the system using the actual steady state of the system as the "initial state" for the transients. This is of course easier said than done.

Two major difficulties present themselves:

1. The severe computational requirements associated with the solution of large sets of differential equations in short time.
2. The sensitivity of the computed results to parameter values.

Much work needs to be done to determine feasibility. It is felt that considerable reduction in complexity can be achieved by simplifying the network "equivalencing", i.e. finding simple models for those portions of the system which are in some sense far removed from the fault.

Transient instability is probably the greatest potential operating danger and improved means for continuous assessment of the stability margins must be deemed as highly desirable. This area is worthy of sustained research efforts.

Dynamic Energy Balance

Closely related to the problem of dynamic stability is the problem of frequency transients. In a typical dynamic stability run one integrates the swing equations over about one or two seconds to determine whether the system is ripping apart.

Even if the synchronizing glue will keep the system together during these first instants the danger may not be past. Groups of generators constituting "islands" in the overall system may jointly be losing frequency over periods of the order of minutes. To prevent the frequency from reaching dangerously low values (< 58 cps) one resorts to methods like load shedding to restore the speed. The operator would be greatly aided in his work if he could simulate in real time these frequency transients, and particularly how they would respond to various load shedding strategies.

The term "dynamic energy balance" has recently(4) been coined to describe this type of analysis that concentrates on frequency rather than rotor angle dynamics. Without going into the details of explaining how this type of analysis differs from dynamic stability analysis we conclude that the results published so far indicate that the computational efforts do not exceed those required for dynamic stability studies. It seems entirely feasible that future DC operators shall have this additional tool available to them.

The Computer As Part Of The Protective Relay System

Historically the protective relay has played a very important part of power systems control. The relays are the "silent sentinels" designed to perform very specific sensing and actuating jobs with the intent of both protecting the system hardware from damage and also, maintain maximum system integrity during disturbances.

Many relays are working as on-off devices with some built-in simple logic ability. Performing logic operations and working on on-off signals is, of course, just what digital computers are best at. It is, therefore, quite reasonable to ask the question whether the digital computer may do some of the jobs which now are reserved for relays and maybe even do them better.

Several such suggestions have been recently made. For example, it has been suggested to use mini-computers in substations to implement relays for various zones and functions.(2) One of the predicted advantages of computers over relays is their better reliability. Also the computer can handle far more complex logic situations than relays and, most importantly, do the fault evaluation much faster.

Another use of the computer in conjunction with relays has been suggested by DyLiacco.(5) A classic difficulty in system operation is diagnosis of what goes on in a system during a cascading fault when several breakers are involved and the fault results in complex changes of system structure. The operator, in order to be able to take proper corrective commands must make a fast assessment of the situation. It is possible to develop computer programs which will present instantaneous assessments of the actual situation and thus save time for the operator to take vital actions.

SUMMARY

Control of a modern power system is a many-faceted problem complex encompassing all operational actions which have as their final objective to maintain the system in its normal operating state.

System complexity dictates that the control system designer must always subdivide the overall system into smaller, but manageable, pieces.

The human operator will have an increasingly important role in the control centers of the future, where his unmatched decision making ability never will be replaced but certainly enhanced by computational robots.

REFERENCES

1. O. I. Elgerd, "Electric Energy Systems Theory: An Introduction," (book), McGraw-Hill Book Co., 1971.

2. "Power System Computer Feasibility Study," Report in 4 volumes from IBM Research Division, San Jose, Calif., Dec. 1968.

3. T. E. DyLiacco, "The Adaptive Reliability Control System," IEEE Transactions on PES, Vol. 86, #5, May 1967, pp. 517-528.

4. K. N. Stanton, "Dynamic Energy Balance Studies For Simulation of Power-Frequency Transients," 1971 PICA Proceedings, pp. 173-179.

5. T. E. DyLiacco, "Processing by Logic Programming of Circuit-Breaker and Protective-Relaying Information," IEEE Transactions on PES, Vol. 88, Feb. 1969, pp. 171-175.

TRENDS IN STABILITY ANALYSIS OF POWER SYSTEMS

A. A. FOUAD
Electrical Engineering Department
Iowa State University
Ames, Iowa 50010

INTRODUCTION

The subject of power system stability is concerned with the behavior of the synchronous machines in the system after they have been perturbed. The problem has been with us since the 1920's, but has received particular attention since the 1960's. Two phenomenae have recently been experienced with rather serious consequences. Certain severe impacts occurred in some electrical power systems in the U. S. and Canada which initiated a chain-reaction-type effect resulting in large scale "blackouts". Also several situations were encountered where neighboring electrical power networks could not operate satisfactorily in parallel. Growing oscillations, under seemingly normal operating conditions, forced the opening of the tie-lines between the systems.

To overcome the above problems, the operating utilities, the power pools, and the regional coordinating councils have adopted certain measures designed to improve the security of power systems. In addition, greater emphasis has been given to the study of the dynamic behavior of the power system. Some aspects of this dynamic behavior will be covered by other speakers in this session. This presentation will be concerned only with the performance of a power system when subjected to impacts, particularly major or severe impacts.

In this paper I intend to give a brief description of the problem of power system stability, and a historical perspective of the changes in our view of the problem. The needs dictated by these factors are discussed, and the main directions which the research in this field has taken are outlined.

Since the subject is too broad to be covered in detail in one presentation, an implied value judgement has been made by this writer as to what is significant and what is a representative sample of an effort in a given area. Also an attempt will be made to survey the field in general and in some cases overlook some problems which might be of regional interest or importance. It is hoped that the reader will be stimulated to make his own investigation of the subject.

POWER SYSTEM STABILITY

As previously mentioned, the problem of power system stability is concerned with the ability of the synchronous machines to remain in synchronism after they have been subjected to any impact or perturbation. If the perturbation does not involve any net change in power, the machines should return to their original operating state. If an unbalance between the supply and demand is created due to a change in load, in generation, or in the network conditions, a new operating state is necessary. For "stable" operation all synchronous machines should remain in synchronism. In addition, the oscillatory transient following a disturbance must be damped out and the system settles to a new steady operating state in a finite time.

Distinction is made in the literature between the situation when the power system, and hence the synchronous machines, is subjected to large impacts, and that when smaller and more normal impacts occur. The former is referred to in the literature as "transient stability", and the latter is known as "dynamic stability". For successful planning and operation of a power system, both questions of transient and dynamic stability must be answered to the satisfaction of the power system engineer. If in the transient following an impact, the rotors of the synchronous machines "swing" far enough so that one or more machines lose synchronism with the rest of the machines in the system, this system is unstable. However, if during a transient the rotor angles, measured with respect to a synchronously rotating reference, maintain relatively small angle differences such that synchronism is not lost, the system is considered stable.

The usual method of conducting a power system stability study is to make a mathematical model of the system during the transient[1,2]. The elements included in the model are those which affect the acceleration (or deceleration) of the machine rotors. These elements included in the model are: the network, the loads, the synchronous machines and their control equipment, the prime-movers and their control equipment, and other supplementary controllers if necessary.

The equations describing a power system in a transient state are highly nonlinear. In dynamic stability studies, these equations are usually linearized, and conventional linear system analysis techniques are used for analysis. In transient stability studies, time solutions of the machine rotor angles is the normal method of analysis.

HISTORICAL PERSPECTIVE

The Situation in the 1950's

Until late in the 1950's the power systems in the North American continent operated in small groups or small power pools. Interconnections on the scale known today, forming very large integrated systems, were not common. The synchronous machines operating in parallel were "electrically close", and the ties connecting them were "electrically strong". The transient behavior of such a system was dominated by the inertia effects and the synchronizing forces holding the machines together. This system would be analogous to a group of masses held together with stiff springs. Each mass, when perturbed, would have a dominant mode of oscillation. Different modes of oscillations would not likely reinforce each other to cause amplified delayed oscillations. Also such a system would have many inherent sources of damping.

A power system thus described had the additional features of relatively slow, and by today's standards, rather unsophisticated control equipment. To determine whether synchronism is maintained in a transient, the engineer needed only to examine the first oscillation or swing for all machines. The period of interest in such a study would be in the order of one second or less.

The above considerations, together with the tools available to the engineer at the time, led to the adoption of a very simplified mathematical model for the power system in a stability study. The basic questions to be answered were usually of the form: will synchronism be maintained if the system is subjected to a given fault cleared at a specified time? What is the critical clearing time for such a fault?

We should point out here that the mathematical models used, and the tools available then to solve them provided the answers needed by the engineer. If in the design or planning stage a stability problem was anticipated, solutions were provided in the form of faster breakers, improved network layout, stronger ties, etc.

New Factors to be Considered

By the 1950's the power systems in North America have changed significantly such that in many cases the situation described above was no longer valid. Some of the new factors to be considered:

1. Large scale interconnections to form large integrated power systems have become commonplace. The power systems in North America form two main interconnected networks of enormous size. There are many operating advantages to such interconnections, but from a stability point of view they bring with them the following problems:
    a. They extend over large distances and connect generators which are electrically far apart; the synchronizing forces among these generators are weakened.
    b. Since in the course of normal operation the tie-lines between power pools or system may carry sizable amount of power, the tie-lines have acquired great significance; the loss of a tie is similar to the loss of generation and may have the same consequence.
    c. Some impacts may isolate whole systems or regions, if the ties are severed, and the effect may cascade into large scale black outs.
    d. Situations are encountered where objectionable oscillations are observed on "weak" tie-lines.
2. In the situation described above some rather "new" phenomenae were observed:
    a. During some transients, the synchronous machines survived the first swing, but synchronism was lost at a later swing.
    b. The phenomena of growing oscillations observed on some tie-lines is indicative of the presence of negative damping in the system.
    c. There has been cases where analysis, based on a certain mathematical model, predicted stability and in practice they proved to be unstable and vice versa.
3. Modern control equipment is much faster and is more sophisticated than that previously available. If they are not adjusted properly they may contribute to instability. On the other hand, if properly coordinated with the other system components they may help improve the dynamic system performance.

The Needs

The factors outlined above have dictated that the engineer reexamine his view of the power system stability problem. Not only were there new factors to be considered and new phenomenae to be explained and analyzed, but also there was evidence that the available methods were in some cases deficient and suspect. By the mid 1960's it has become apparent that the following was needed:

1. To improve upon our understanding of the dynamic behavior of the power system. This would be helpful in both predicting its transient behavior and in correcting "troubles" when they occur.

2. To develop a more accurate mathematical model so that its dynamic behavior can be predicted more accurately.

3. To extend the period of the transient time to be analyzed. This means that some effects, previously neglected, must be considered; and more components should be incorporated in the model.

4. Develop better tools of analysis and computation of the system dynamic behavior.

5. All aspects of the system stability should be examined: loss of synchronism transients; loss of generation (or ties) transients; and dynamic behavior.

RESEARCH EFFORTS AND TRENDS

A. System Modeling

Considerable work has been reported in the literature on the mathematical modeling of the power system components.

1. Modeling of the synchronous machine:

The mathematical model of the synchronous machine given by Park and others in the 1920's was considered too complicated for use in most stability studies. It has been perhaps somewhat in a modified form reintroduced[3], and it is now a common practice to represent some machines in the system in greater details than was previously possible. In the late 1960's state-space formulation of the machine equations was adopted[4]. The dynamic model of the machine, the number of circuits to be used in the model, and the parameters of these circuits have received some attention recently and efforts are under way to determine the effect of the machine model on the system dynamic behavior[5,6]. Thus current research efforts involve both the development of adequate models and the "philosophy of modeling". It is now recognized that the accuracy of the results obtained in a stability study depends on the type of study, on the complexity of the machine models used, and on the parameters used in these models. At the same time increasing the complexity of the model does not necessarily give better results unless accurate data and parameters are used.

2. Modeling of Control Equipment

U. S. manufacturers have agreed to a common format for computer representation of excitation systems in stability studies[7,8].

Commercially available excitation systems are grouped into main categories which can be represented by the same mathematical model or block diagram. Typical parameters are made available to the users. Experience with this type of representation has been satisfactory except for the fact that some users found discrepancy in the published data on excitation system parameters[9]. More recently some users have expressed interest in broadening the list of excitation system types to include non-U. S. made systems which do not fit the previously published list.

For modeling the mechanical turbine and governor some block diagrams are commonly used in most available stability computer programs. These models usually take into account the time lags introduced by the servomotor, the steam bowl and reheat circuit in steam plants. They in effect consider the steam supply to be inexhaustible. If the stability study is to be conducted for a transient period in the order of few seconds this may not be grossly in error. For longer periods a model of the boiler and other power plant components should be included. Considerable work is currently being done to develop such models[10,11].

In extended stability studies, it is felt that supplementary controllers, such as the load-frequency controllers, should be included in the system's mathematical model.

3. Modeling of loads

In the past loads were represented in stability studies by constant impedance, determined from pretransient conditions. It is now generally agreed that this model is highly suspect. Some emphasis has been given to more accurate representation of the system loads during transients. The difficulty lies in the fact that very little data was available in the literature concerning the transient behavior of composite loads. It is generally recognized that they are both voltage and frequency sensitive. Some recent work has been devoted to the determination of the load constants[12].

4. Modeling of Other System Components

As explained above, modeling of the main components of a power system have received considerable attention. Additional work is needed to determine the adequacy of the models used for such elements as relays, on load tap-changing transformers, automatic generation control and others.

Some attention has been given to the modeling and analysis of high voltage d.c. systems[13].

B. The Dynamic Stability Problem

In the 1960's it became evident to the power system engineers that the natural damping in large interconnected systems was frequently too weak to prevent instability. This was experienced in situations where growing oscillations were encountered under "normal" operating conditions, and in few cases where transient instability occurred after the first swing. This problem received considerable attention in the late 1960's. It was recognized that the system damping could be improved by the proper adjustment of the excitation system parameters and by adding certain auxiliary stabilizing signals to the excitation. The basic scheme usually involves a signal from frequency or speed deviation, processed and fed to the excitation system. The processing network is usually a network the transfer function of which provides the proper phase relation to give improved torque at the desired frequency [14,15].

The use of excitation control to improve the dynamic behavior of the power system has become widely accepted in the industry[16-18]. To a large extent this problem seems to have been brought under control.

An interesting application of the same techniques used in this type of analysis is to determine the behavior of a power system when subjected to "potentially troublesome" loads, such as a large cyclic load produced by a nuclear accelerator[19].

C. Computational Techniques

In a transient stability study involving many synchronous machines and loads, the mathematical model becomes exceedingly complex. Synchronous machine equations are given to their own reference frame, and these equations are "tied" together by the power network. In obtaining time solutions numerically, significant problems are encountered in solving the different sets of equations simultaneously. This is known in the literature as the "interface problem". Furthermore the sheer size of the model makes the time solutions quite costly. Also the results obtained, unless they are conveniently displayed and presented, may be rather difficult to analyze. The computational techniques have received the attention of the researchers in all the above areas. Research is being conducted by many on numerical integration algorithms to improve their efficiency, to reduce interface errors, and to reduce the computer time needed for obtaining a solution[20-27]. Among the most promising schemes are: the fast transient solutions[20] claimed to reduce the computer time by an order of magnitude, and the dynamic energy balance scheme where the transient is separated into different types dominated by basically different parameters and time constants[26].

Concerning the results, commercially available computer programs have certain features for convenient display of results such as tabulated output, plots of selected parameters versus time, storing of output data on microfilm, etc. Perhaps a step forward in interpreting the results obtained in a stability study is the method of identifying groups of machines which swing coherently, and the synchronizing capabilities of transmission networks[28].

D. Selective Simplifications of Large Systems

A detailed mathematical description of one synchronous machine with its control equipment may require more than fifteen nonlinear differential equations plus a set of additional auxiliary algebraic equations. A large power system may comprise tens or even hundreds of these machines. The need to adopt the simplest mathematical model which gives an adequate answer for the problem at hand is quite evident. Two basic approaches are adopted: 1- selection of simplified models of system components and 2- use of dynamic equivalents which reduce the mathematical size of the system and the order of equations. In the former the identity of the system components are retained while in the latter no effort is made to maintain the identity of the elements of the system being equivalenced.

The question of the region of a network to be represented by an equivalent is usually decided by the engineer's judgement or personal knowledge of the power system. An effort to establish, in a stability study, distance measures and coherence recognition of the boundaries of the areas to be retained and those to be reduced has recently been suggested[29].

Dynamic Equivalents

Originally the basic method of making a dynamic equivalent of an area containing synchronous generators was to obtain a network equivalent and assign the proper inertia to it. This preserved the influence of the reduced region on the dominant mode of oscillation for each of the machines to be retained. Some techniques which were recently suggested are basically mere refinements of this basic concept[30,31]. The following recent efforts have more theoretical backing:

> A technique to give an equivalent of the governor-turbine dynamics on the average system frequency has recently been offered[32]. This is similar (conceptually) to the dynamic energy balance idea previously discussed[26].
>
> A rigorous technique for reducing the system size of an "outside area" was suggested. It is based upon obtaining, at the nodes of interconnection between system to be retained and that to be reduced, a set of linearized equations in state-space form. These equations are transformed into canonical form. The order of the system is then reduced by a systematic procedure based on the nature of the eigenvalues of the system and the nature of the coupling elements in the transformed matrices[33,34]. Thus modes which decay fast and modes which are weakly coupled to the rest of the system are deleted.

E. Theoretical Investigation of Stability

Considerable effort has been devoted to the subject of determining the stability of power systems via direct methods (in contrast to obtaining time solutions). The method which received considerable attention is Liapunov's second method. The basic concept is to construct a so-called Liapunov function which possesses certain properties in a specified region of state-space. Liapunov functions constructed for power systems are energy integrals or approximate energy integrals[35,36]. Reference[37] investigates 14 possible functions for a medium size power system. The region of stability is determined by the stable and unstable equilibrium states of the system. Stability is determined by knowing whether the final equilibrium state (after the transient) falls inside or outside the region of stability.

The status of this line of investigation is that it has been limited to the most elementary mathematical model of a power system. Since the construction of the Liapunov functions has been linked to the system energy, no account has been taken for higher order models. Even for the elementary model the results obtained have been frequently too conservative. In addition, some problems were encountered with identifying the transient energy which has the greatest effect on transient stability, and with the characterization of the equilibrium and nonequilibrium states of the system. In earlier works, these matters were decided more or less by intuition. Some recent work has shed some light on these subjects[38,39].

A theoretical problem related to the determination of stability by direct methods is the subject of stability indices. Sensitivity analysis has been used to investigate the margin of stability in transient studies[40], and more recently pattern recognition has been proposed as a means to determine transient security[41].

Finally, optimal control theory is being suggested for stabilization of multimachine power systems[42]. This is primarily offered as an aid in the dynamic stability problem.

F. Self Excited Oscillations

Series capacitors have been used to compensate the reactance of long distance transmission lines. This is deemed desirable to improve the stability of the power systems incorporating these lines Recently it was discovered that certain generators connected to the compensated lines have suffered damage from torsional resonnance. Investigation of the phenomena showed that the combination of series capacitor compensated system and the torsional mechanical modes of a turbine generator may interact in such a way as to introduce instability. Eigenvalue analysis of the combination of electrical and mechanical systems can predict regions of instability and possible remedial methods[43,44].

CONCLUDING REMARKS

The attention given to the problem of power system stability in the last few years was dictated by the presence of emergency or near emergency conditions. At stake was the whole question of whether large integrated systems could continue to operate in the face of such problems as growing oscillations and black outs. Power System Stability analysis was a "cook book" type procedure followed only when new generators are planned, and understood by few.

The bibliography at the end of this paper is but a small fraction of the U. S. technical literature on the subject in the last few years. Many technical sessions, conferences, symposia and discussions were devoted to the subject in the IEEE. The result of all this is a better knowledge of many aspects of this complex subject. What is more important is that this knowledge has been translated into practice to improve system security, and into greater understanding and awareness on the part of the power system engineer.

To illustrate this point we need only to mention that the system oscillation problem, which was a mystery only a decade ago, has been basically solved with improving the system damping artificially with excitation control. Better understanding of the system dynamic behavior is further illustrated by such practices as load-shedding during severe transients involving loss of generation, and by the introduction of generation control schemes such as fast valving[45]. Designing generator and system components with the stability problem in mind is now an accepted practice. Commercially available stability computer programs allow a fairly elaborate representation of the machines, prime movers, loads, and relays. All these are definite improvements and are signs of progress.

While much progress has been achieved, we should not relax our efforts to continue research to enhance our understanding of the problem of power system stability and to improve our tools of analysis. While impressive gains were made, much more remains to be done. To prove my point, I would like to remind the reader that at the time of this writing, the only way to analyze a system comprising a network of 2000 buses, few hundred generators, and a large number of loads is to obtain time solutions. Thus, hundreds of parameters, buried in many more equations, are solved for by the computer. After spending some thousands of dollars, and burdened with stacks of paper output, the engineer will have to answer some basic questions such as: will the system be stable under the given impact? Surely the answer is buried in the stack of output he receives, but I cannot help but feel that there ought to be a better way to do this job.

Bibliography

1. Young, C.C. Equipment and System Modeling for Large Scale Stability Studies. IEEE Trans., Vol. PAS- 1, pp. 99-109.
2. Ramey, D.G., Byerly, R.T.,Dynamic Simulation of Interconnected Systems. IEEE-Power Ind.Comp.Appl.Conf., 5th Conf.Rec. May 15-17 1967, pp.31-40.
3. Recommended Phasor Diagram for Synchronous Machines. Anon, IEEE. Trans Vol. PAS-88, No. 11, 1969, pp. 1593.
4. Undrill, J. M. Structure in the Computation of Power System Dynamical Response. IEEE Trans., Vol. PAS-88, pp. 1-6.
5. Schultz, R.P., Jones, W.D., Ewart, D.N. Dynamic Models of Turbine Generators Derived from Solid Rotor Equivalent Circuits. IEEE Trans. Vol. PAS-92, pp. 926-933.
6. Dandeno, P.L., Hauth, R.L, Schultz, R.P. Effects of Synch.Mach. Modeling in Large Scale System Studies. IEEE Trans.,Vol.PAS-92, pp.574-582.
7. Computer Representation of Excitation Systems. IEEE paper no. 31TP67-424, presented by the IEEE Working Group of the Excitation Systems Subcommittee at the Summer Power Meeting, Portland, Oregon, 1967.
8. Excitation System Dynamic Characteristics - IEEE Committee Report. IEEE Trans., Vol..PAS-92, pp. 64-75.
9. Schleif, F.R. Excitation System Modeling. A paper presented at the symposium on Adequacy and Philosophy of Modeling, IEEE Summer Power Meeting, San Francisco, 1972.
10. Dynamic Models for Steam and Hydro Turbines in Power System Studies. IEEE paper no.T73089-0, presented by IEEE Task Force on Overall Plant Resp., Dynamic Syst.Perf. W.G. at IEEE Winter Power Meeting.
11. Laubli, F., Fenton, F.H. The Flexibility of the Super-critical Boiler as a Partner in Power System Design and Operation: Parts I & II. IEEE Trans., Vol. PAS-90,,pp. 1719-1733.
12. System Load Dynamics-Simulation Effects & Determination of Load Constants. IEEE Comm.Report. IEEE Trans., Vol. PAS-92, pp. 600-610.
13. Carroll, D.P., Krause, P.C. Stability Analysis of a d.c. Power System. IEEE Trans., Vol. PAS-89, No. 6, pp. 1112-1119.
14. DeMello, F.P., Concordia, C. Concepts of Synchronous Machine Stability as Affected by Excitation Control. IEEE Trans.,Vol.PAS-88, pp. 316-329.
15. Schleif, F.R., Hankins, H.D., Hattan, E.E., Gish, W.B. Control of Rotating Exciters for Power System Damping: Pilot Applications and Experience. IEEE Trans., Vol. PAS-88, pp. 1259-1266.
16. Klopfenstein, A. Experience with System Stabilizing Excitation Controls on the Generation of Southern California Edison Co. IEEE Trans. Vol. PAS-90, pp. 698-706.
17. Warchol, E.W., Schleif, F.R., Gish, W.B., Church, J.R. Alinement and Modeling of Hanford Excitation Control for System Damping. IEEE Trans. Vol. PAS-90, pp. 714-724.
18. Bobo, P.O., Skooglung, J.W., Wagner, C.L. Performance of Excitation Systems Under Abnormal Conditions. IEEE Trans.Vol.PAS-87, pp.547-553.
19. Pinnello, J.A., Van Ness, J.E. Dynamic Response of a Large Power System to a Cyclic Load Produced by a Nuclear Accelerator. IEEE Trans. Vol. PAS-90, pp. 1856-1862.
20. Dommel, H.W., Sato, N. Fast Transient Stability Solutions. IEEE Trans. Vol. PAS-91, pp. 1643-1650.

21. Dommel, H.W. Nonlinear and Time-varying Elements in Digital Simulation of Electromagnetic Transients. IEEE Trans.,Vol.PAS-90, pp. 2561-2567.
22. Oswald, R., Smith, O.J.M. Integration of Nonlinear Differential Systems with Wide Eigenvalue Range. IEEE Trans.,Vol. PAS-90, pp. 2586-2589.
23. Talukdar, S.N. Iterative Multistep Methods for Transient Stability Studies. IEEE Trans., Vol. PAS-90, pp. 96-102.
24. Steinbrenner, G., Gelopulos, D.P. An Algorithm for Transient and Dynamic Stability Solutions Without Interface Error. 1973 PICA Conf. pp.113-117.
25. Fuller, H. L., Hirsch, P.M., Lambie, M.B. Variable Integration Step Transient Analysis: VISTA. 1973 PICA Conf. Proceedings, pp. 277-284.
26. Stanton, K.N. Dynamic Energy Balance Studies for Simulation of Power Frequency Transients. IEEE Trans., Vol. PAS-91, pp. 100-117.
27. Part I; Janischewskyj, W., Prabhashankar, K. Part II; Dandeno, P.,Prabhashankar, K. Simulation of the Nonlinear Dynamic Response of Interconnected Synchronous Machines. IEEE Trans.,Vol. PAS-91, pp. 2064-2077.
28. Ramey, D.G., Byerly, R.T., Sherman, D.E. The Application of Transfer Admittances to the Analysis of Power System Stability Studies. IEEE Trans. Vol. PAS-90, pp. 993-999.
29. Lee, S.T.Y., Schweppe, F.C. Distance Measures & Coherency Recognition for Transient Stability Equivalents. IEEE paper no.T73092-4, presented at 1973 Winter Power Meeting, New York, Jan. 1973.
30. Brown, H.E., Shipley, R.B. A Study of Stability Equivalents. IEEE Trans. Vol. PAS-88, pp. 200-207.
31. Chang, A., Adibi, M.M. Power System Dynamic Equivalents. IEEE Trans., Vol. PAS-89, pp. 1737-1744.
32. Chan, M.L., Dunlop, R.D., Schweppe, F. Dynamic Equivalents for Average System Frequency Behavior Following Major Disturbances. IEEE Trans., Vol. PAS-91, pp. 1637-1642.
33. Undrill, J.M., Turner, A.E. Construction of Power System Electromechanical Equivalents by Model Analysis. IEEE Trans.,Vol.PAS-90, pp.2049-2059.
34. Undrill, J.M., Casazza, J.A., Gulachenski, E.M., Kirchmayer, L.K. Electromechanical Equivalents for Use in Power System Stability Studies. IEEE Trans.,Vol. PAS-90, pp. 2060-2071.
35. El-Abiad, A.H., Nagappan, K. Transient Stability Regions of Multimachine Power Systems. IEEE Trans.,Vol. PAS-85, pp. 169-179.
36. Lúders, C.A. Transient Stability of Multimachine Power Systems Via the Direct Method of Lyapunov. IEEE Trans.,Vol. PAS-90, pp. 23-36.
37. Williams, H.F., Louie, S.A., Bills, G.W. Feasibility of Liapunov Functions for the Stability Analysis of Elective Power Systems Having Up to 60 Generators. IEEE Trans.,Vol. PAS-91, pp. 1145-1157.
38. Tavora, C.J., Smith, O.J.M. Characterization of Equilibrium and Stability in Power Systems. IEEE Trans.,Vol. PAS-91, pp. 1127-1130.
39. Tavora, C.J., Smith, O.J.M. Equilibrium Analysis of Power Systems. IEEE Trans.,Vol. PAS-91, pp. 1130-1137.
40. Rahimi, A., Stanton, K.N. Dynamic Aggregation and the Calculation of Transient Stability Indices. IEEE Trans.,Vol. PAS-91, pp. 118-122.
41. Pang, C.K., Prabhakara, F.S., El-Abiad, A.H., Koivo, A.J. Security Evaluation in Power Systems Using Pattern Recognition. IEEE paper T73107-0, presented at the Winter Power Meeting, New York, Jan. 1973.
42. Yu, Y., Moussa, H.A.M. Optimal Stabilization of a Multimachine System. IEEE Trans., Vol. PAS-91, pp. 1174-1182.
43. Ballance, J.W., Goldberg, S. Subsynchronous Resonnance in Series Compensated Transmission Lines. IEEE paper T73167-4 presented at Winter Power Meeting, New York, January 1973.
44. Bowler, C.E.J., Ewart, D.N., Concordia, C. Self Excited Torsional Frequency Oscillations with Series Capacitors. IEEE paper T73218-5 presented at the Winter Power Meeting, New York, Jan. 1973.
45. Cushing, Drechster, Kilgor, Marshall, Stewart. Fast Valving as an Aid to Power System Transient Stability. IEEE Trans.,Vol.PAS-91,pp.1624-1636.

DATA BASE
FOR
REAL-TIME POWER SYSTEM CONTROL*

by

A. S. Debs
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

Modern real-time power system control is dependent on an elaborate data base consisting of hundreds of sensors located all over the power system, telemetry equipment, and computers to process the on-line data and perform the functions of state estimation and parameter identification. An overview of the state-of-the-art in this area is presented with emphasis on computer methods employed and the tradeoffs involved in data base design.

INTRODUCTION

At present, several major power utilities, are in the process of constructing or designing advanced energy control centers for real-time computerized power system control. Obviously each utility has its own philosophy as well as its individual problems. The roles these energy control centers are supposed to play may vary but can be generalized into two basic ones. First, there is the job of replacing old and existing supervisory dispatch equipment with central and local digital systems. Under this category falls such important control functions like automatic generation control and economy dispatch. Second, there are the areas of advanced control computer applications. Notable among these are security monitoring, transient stability monitoring, optimum power flow (an advanced version of economy dispatch), corrective rescheduling, voltage control, and others. It is to this second set of areas that considerable attention has taken place in the past six years or so.

The success or failure of the advanced applications areas depends on many factors. An important factor is their reliance on an elaborate on-line data base which can provide an updated picture, every few seconds, of the system's operating conditions, network status as well as its coupling with other neighboring systems.

The type of data obtained is duscussed in the next section. Follwoing that, the techniques of state estimation are discussed. Techniques of rejecting excessively noisy data are then discussed. The utilization of online data collected over a long-time period in estimating system parameters is elaborated upon. Finally, a discussion of the factors influencing the location and number of sensors used as they affect overall data base design is provided. The conclusion section contains some suggestions for future applications.

---

*Part of the research effort reported in this paper is sponsored by the National Science Foundation under grant No. GK-37474, with the endorsement and technical support of the Bonneville Power Administration, Portland, Oregon.

REAL-TIME DATA

In Figure 1 a typical portion of the "main-grid" high-voltage network of the Bonneville Power Administration is shown. Indicated on the figure are locations and types of quantities that are measured. The types of measurements used are

- KV at selected busses (nodes)
- MW and MVAR at the ends of many transmission lines
- Net MW and MVAR flowing into a bus from the external network. This net "injection" is equal to the sum of line flows out of the bus.

Additional quantities which are not shown are also measured. For example, breaker status for each breaker is telemetered. Current magnitudes are measured mostly for protective relaying.

All the analog data is digitized and telemetered once every few seconds to the control center. In Figure 2 we show a block diagram which describes the various functions involved in the utilization of such data. Breaker status data is used to update the configuration of the telemetered network. An anomaly detection and diagnosis program analyses the measurement data and interacts with the state estimator program in order to check the validity of all the data, identify "bad data" and reject it. The state estimation program will receive the validated data and perform a least-squares estimation of complex voltage variables (state variables) at all network nodes. In turn, these are used to calculate all power flow and injection quantities in the network. These state and flow variables form the data base for security monitoring, corrective rescheduling and the other advanced applications functions.

THE STATE ESTIMATOR PROGRAM

The state estimator program is the main computational tool needed to calculate various system variables. By understanding it, other problems and issues of data base design become easier to handle.

Problem Description

Let z denote the vector of measurements obtained by a particular scan of the telemetry system, and let x denote the vector of voltage magnitudes and phase angles at all network busses. The relationship between x and z is given by

$$z = h(x) + v \tag{1}$$

where $h(\cdot)$ describes the non-linear algebraic relationship between x and the true values of measured quantities; and v is a vector representing measurement error. The main assumption in (1) is that v is a zero-mean Gaussian random vector with a known (diagonal) covariance matrix

$$R = E[v\ v^T]. \tag{2}$$

Furthermore, the measurements are assumed to be redundant in the sense that $\dim[z] = m \geq \dim[z] = n$.

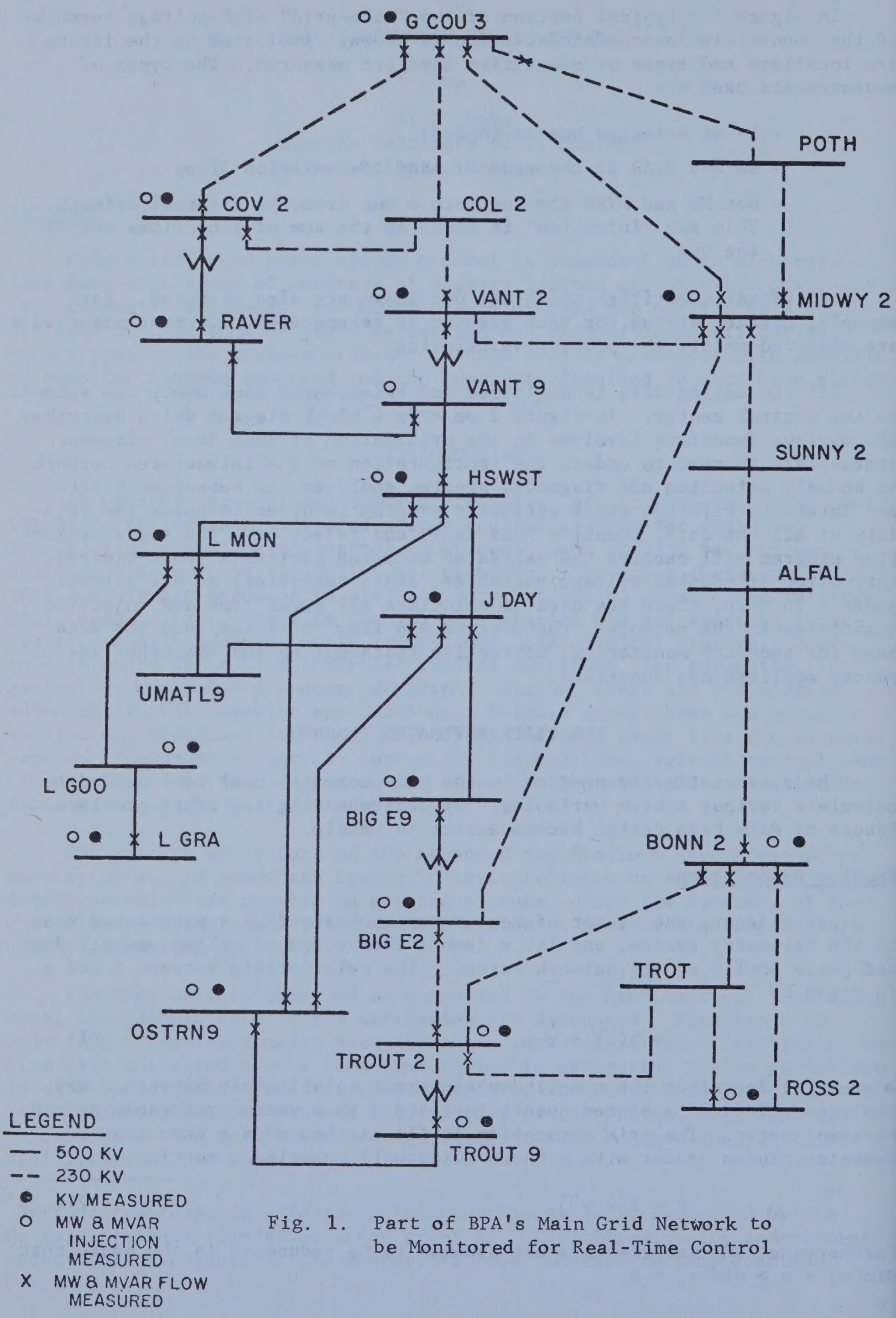

G COU 3
POTH
COV 2
COL 2
VANT 2
MIDWY 2
RAVER
VANT 9
SUNNY 2
HSWST
L MON
ALFAL
J DAY
UMATL9
L GOO
L GRA
BIG E9
BONN 2
BIG E2
TROT
OSTRN9
TROUT 2
ROSS 2
TROUT 9
LEGEND
— 500 KV
-- 230 KV
● KV MEASURED
○ MW & MVAR INJECTION MEASURED
X MW & MVAR FLOW MEASURED

Fig. 1. Part of BPA's Main Grid Network to be Monitored for Real-Time Control

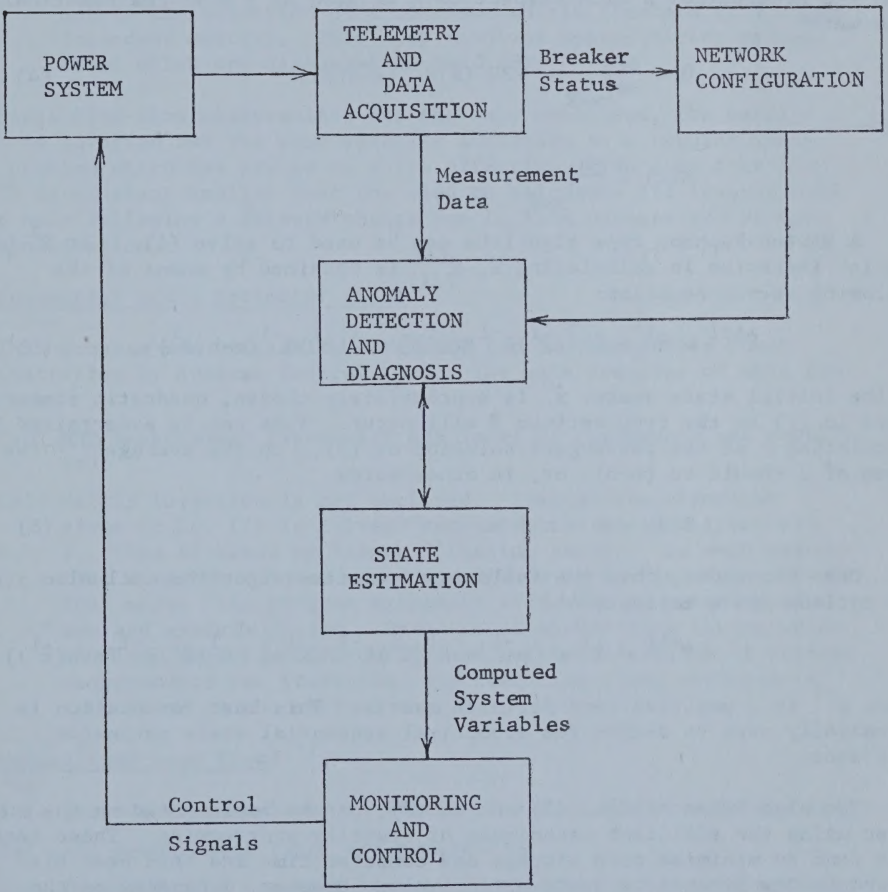

POWER SYSTEM
TELEMETRY AND DATA ACQUISITION
Breaker Status
NETWORK CONFIGURATION
Measurement Data
ANOMALY DETECTION AND DIAGNOSIS
STATE ESTIMATION
Computed System Variables
Control Signals
MONITORING AND CONTROL

Fig. 2. Role of Data Collection and Processing in On-Line Power System Control

Since there are more equations than unknowns, a least-squares estimate $\hat{x}$ which minimizes

$$J = (z - h(\hat{x}))^T R^{-1} (z - h(\hat{x})) \tag{3}$$

is sought.

Solution Methods

Several solution methods to obtain the minimum of J are available. They can be classified into exact and approximate methods.

In the exact methods, numerical algorithms are developed to ascertain that the gradient of J with respect to x is zero at $x = \hat{x}$. Mathematically this means

$$0 = \left.\frac{\partial J}{\partial x}\right|_{x=\hat{x}} = -2H^T(\hat{x})R^{-1}(z-h(\hat{x})) \tag{4}$$

where

$$H(\hat{x}) \triangleq \left.\frac{\partial h}{\partial x}\right|_{x=\hat{x}} .$$

A Newton-Raphson type algorithm can be used to solve (4). Let $\hat{x}^i$ be the ith iteration in calculating $\hat{x}$, $\hat{x}^{i+1}$ is obtained by means of the following vector equation:

$$\hat{x}^{i+1} = \hat{x}^i + [H^T(\hat{x}^i)R^{-1}H(\hat{x}^i)]^{-1} H^T(\hat{x}^i)R^{-1}(z-h(\hat{x}^i)). \tag{5}$$

If the initial state vector $\hat{x}^o$ is appropriately chosen, quadratic convergence in (5) to the true optimum $\hat{x}$ will occur. This can be ascertained by calculating J at the convergent solution of (5). On the average$^{(23)}$ the value of J should be (m-n), or, in other words

$$E(J) = m - n > m. \tag{6}$$

Debs has shown, that the following iterative algorithm will also yield the optimum state estimate$^{(11)}$

$$\hat{x}^{i+1} = \hat{x}^i + [p^{-1} + H^T(\hat{x}^i)R^{-1}H(\hat{x}^i)]^{-1} H^T(\hat{x}^i)R^{-1}(z-h(\hat{x}^i)) \tag{7}$$

where p^{-1} is a positive semi-definite matrix. This last formulation is essentially used to derive the suboptimal sequential state estimator equations.

The algorithms of Eqs. (5) and/or (7), can be implemented on the computer using the efficient techniques of sparsity programming. These techniques tend to minimize core storage and computer time and have been discussed in the literature quite extensively. However, depending on the particular measurement system used the effectiveness of sparsity techniques can be enhanced or limited. Normally, more measurements mean less sparsity and vice-versa.

The approximate solution techniques are designed to provide a sufficiently accurate state estimate with attractive computational characteristics. In what follows in this section, three popular techniques will be discussed.

1. AEP Algorithm(8)

This algorithm was developed by the staff of American Electric Power Service Corporation and is now being used at the AEP Control Center. It is based on two significant points:

(a) Line power flow measurements with one voltage magnitude measurement are the only measurement types considered. This excludes net injection measurements as well as several voltage measurements. Typically, every main-grid transmission lines is measured at both ends.

(b) By means of non-linear transformations the performance index J is modified so that the resulting iterative algorithm requires the inversion of a constant matrix (instead of a state-dependent matrix). This step involves approximating assumptions which are discussed in ref. (8).

Since line-flow measurements are the only ones used, the matrix which is inverted has the same sparsity structure as a regular load-flow problem which has proved to quite effective. The fact that this matrix is constant implies that the need to calculate its inverse will arise only following a network change due to line outages and/or bus sectionalization.

2. Sequential State Estimator (SSE)(11)

This program was originally developed for the Bonneville Power Administration by Systems Control, Inc. The main features of this program are:

(a) All measurement types with any level of redundancy are permitted.

(b) Matrix inversion is not employed. Instead the algorithm given in Eq. (7) is solved "sequentially" at each iteration i. This is based on Kalman Filtering theory. At each sequential step the updated state covariance matrix is diagonalized. This makes the program extremely efficient in terms of storage and execution time. By means of appropriate tuning which involves measurement ordering and multiple scanning of various measurements per iteration, the resulting state estimate is very close to the optimal one.

3. Generalized Load Flow(12)

This program was developed in Italy. The main characteristics are:

(a) All measurement types are admissible.

(b) By means of a constant linear transformation M, a new measurement vector

$$z_{new} = Mz$$

is obtained. z_{new} has the same dimension as the state vector. Hence, load-flow type program is directly applicable.

Numerically, this has the same characteristics as a load-flow program. Deviation from the optimum results from the simplification involved.

REJECTION OF BAD DATA

Least-squares estimation is known to perform adequately only when the errors in the measurements are small and within the limits provided by their statistical distributions. The possibility of highly erroneous measurements, however, exist. These can result from instrument failures, communication noise, inappropriate error detection codes and the like. If such data is permitted in the estimation process, a smearing of the error will occur all over the system and highly inadequate state estimates will result. The effect of this is to reduce the reliability of the data base considerably.

A few techniques have been developed for the purpose of rejecting bad data. In the remainder of this section a description of three of these techniques will be provided.

1. AEP Method(14)

The AEP bad data rejection method is based on the following:

(a) Rejection of bad data at time t is based on collected information at time t only, i.e., this is a static evaluation process. Hence, the capability of the bad data rejection algorithm is strongly related to the degree of redundancy in the system.

(b) Let z_i and $\hat{z}_i$ be the ith scalar measurement and its calculated estimate respectively, and let ρ_i^2 be the calculated variance of $\hat{z}_i$*. If z_i is a "bad data point," then the quantity

$$(z_i - \hat{z}_i)^2/\rho_i^2$$

is assumed to be larger than any other

$$(z_j - \hat{z}_j)^2/\rho_j^2 \ , \ i \neq j$$

This is more or less a practical assumption since no theoretical proof of this is available.

(c) The way bad data rejection is performed is as follows

(i) The measurement vector z is used to obtain state estimate $\hat{x}$ and a calculated measurement vector $\hat{z}$.

(ii) The performance index

$$J = \frac{1}{m}(z-\hat{z})^T R^{-1}(z-\hat{z})$$

is calculated. If J is of the order of $\frac{m-n}{m}$, no bad data exist. However, if J is considerably greater than $\frac{m-n}{m}$, then there is a problem with bad data.

(iii) In case it is concluded in (ii) that bad data exist the quantities

$$s_i = \frac{(z_i - z_i)^2}{\rho_i^2}$$

---

*This is different from the variance of the error in z_i.

are computed. The measurement with the largest s_i is discarded and the procedure is repeated starting with step (i).

Experimental tests, according to the authors of this method, have demonstrated the effectiveness of the method, even in cases where measurement redundancy is low.

2. BPA Method(7)

This method was developed for BPA and is based on the following

(a) Under normal operating conditions small changes in the measured quantities will occur from one time sample to the next. This "time" factor can be incorporated to provide information on bad data.

(b) Detection of whether bad data exist or not is done in a manner similar to the AEP method.

(c) Diagnosis of bad data is done by calculating the quantities

$$s_i = |\hat{z}_i(k) - z_i(k+1)| \ , \ i = 1, \ldots, m$$

where $\hat{z}_i(k)$ is the estimate of $z_i(k)$ and $z_i(k+1)$ is a new measurement of the same ith quantity at time k+1. If any s_i is larger than its expected statistical threshold it is flagged for further analysis. Consequently, simple correlation analysis is conducted among all flagged quantities to determine which ones correspond to bad data.

This method was tested using actual system data for a small system and the results are reported in ref. (7).

3. Merrill-Schweppe Method(13)

This is an interesting method for what the author's call "bad data suppression." It involves the following:

(a) Static estimation is assumed (same as (a) above in the AEP method).

(b) The performance index to be minimized is quadratic near the origin and "square-root" outside a certain region around the origin.

(c) By minimizing the above performance index little weight is given to measurements that have excessive errors. The result is an automatic filtering of the effect of these measurements on the state estimator.

As we learn more about the real-time system we can expect improvements in all of these methods. It is quite conceivable that a realistic combination of the static and time-dependent methods may provide the ultimate answer.

PARAMETER ESTIMATION

Sensitivity analysis studies(20) have shown that inaccuracies in the network parameters, as well as large biases in the measurements, can result in state estimator degredation to varying degrees. This degredation can be

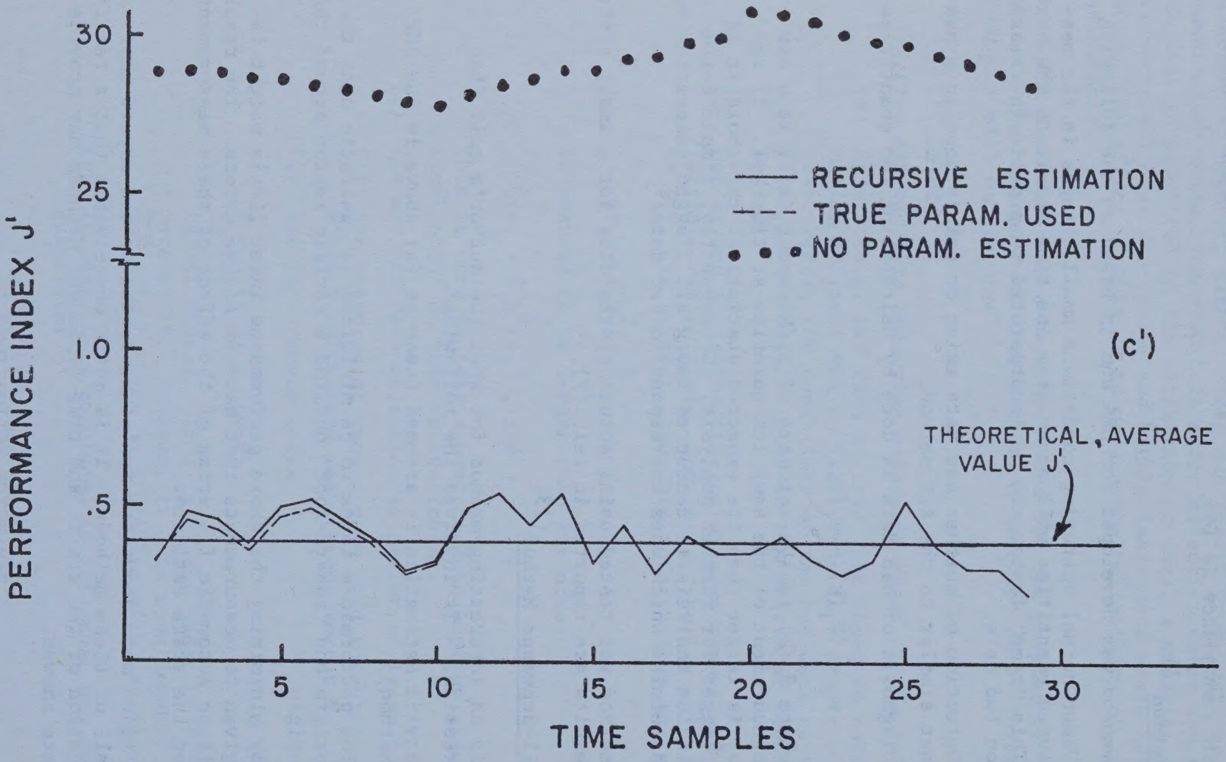


Fig. 3. Comparison of State Estimation Performance Levels where Parameter Errors are Present (Ref. 18).

quite serious in certain cases and defeats the entire purpose of state estimation (see Figure 3).

It has been concluded in two separate studies that parameter errors are indeed present. The first is a German study(16) in which it was concluded that 4-5% errors in the admittance parameters of transmission lines and transformers can be expected. The second study was conducted by Debs and Litzenberger(7) where raw data was used for state estimation. In this study large biases in the measurements were observed. Even when the biases were filtered out, the validity of the network parameter model was still in doubt.

These facts have given rise to the need to tune and/or calibrate the network model to be used in on-line estimation. One way of doing this is to analyze the data using engineering insight and hard manual work. The other, is to develop parameter estimation algorithms to process large amounts of data and estimate the parameters accurately. The feeling of this writer is that both approaches will be needed and should complement each other.

Parameter estimation is done by applying the theory of extended Kalman filtering. The feasibility of this has been demonstrated recently by this writer(18). Future research and development effort will concentrate on approaches and algorithms to efficiently calibrate the parameter model and possibly estimate the parameters of the unobserved external system(18).

DATA BASE SELECTION

The problem of data base selection involves a set of decisions on where, what type, and how many measuring devices should be employed by a particular utility for on-line control. The first thing to note here is that state estimation is a major but by no means the only computational tool that processes the on-line data. Functions such as supervisory control, automatic generation control, billing, protective relaying and others will utilize on-line information which is shared to varying degrees with information needed for advanced application programs.

At present, there is no established theory or practical procedure for data-base selection. Instead, engineering judgement is employed to a great extent. In my opinion the factors that influence this judgement are the following:

1. Definition of the "main-grid" network: This requires a decision on those voltage levels in the power network which will be monitored. For example, AEP only monitors its 765KV network while BPA monitors its 500KV, 230KV, and some 155KV lines. The main factor in this network selection has to do with expected operating conditions and system security.

2. Availability of measuring instruments at existing substations: Usually many, but not all, transmission lines are equipped with potential and current transformers which make it easy to measure KV, MW, and MVAR quantities. These transformers are usually much more expensive than transducers and analog-to-digital converters which will be needed for data transmission. Only a minimal amount of additional instrumentation is justified from an economical viewpoint.

3. Redundancy-reliability tradeoffs: In the opinion of this writer, the main reason for redundancy in the measurement system is not to increase

accuracy of estimated quantities but rather to enhance bad data rejection capability. A very redundant system is reliable and expensive. A non-redundant system is cheap and unreliable. Somewhere, a tradeoff is to be found. Furthermore, the redundancy spoken of in many cases has been a static one. Little concern about the system's evolution in time has been shown. More attention should be payed to "time-redundancy" to enhance overall data base reliability.

4. Computational aspects: There are two kinds of computational problems involved in data base selection. The first is that of redundancy-reliability tradeoff. In the algorithms described above it is safe to say that in all of them computer storage requirements as well as solution execution times are linearly related to the number of measurements. Hence, more measurements mean additional computational requirements. The second is that of computer algorithm selection. For example the AEP algorithm is designed for specific types of measurements, while the BPA algorithm can handle all measurement types. In the opinion of this writer, it is preferrable to retain flexibility in the selection of the algorithm provided that computational requirements do not become excessive.

5. Sensitivity to parameter errors: This area is gaining more and more attention. By selecting the measurement system to be sensitive to parameter errors, parameter estimation can be employed to improve the network model.

Obviously a good deal of analytical as well as experimental work is needed in order to clarify and systematize the selection of the instrumentation system.

CONCLUSION

Significant developments have taken place in the areas of state estimation and rejection of excessively noisy data for on-line control and monitoring of power systems. A few areas of investigation are just starting underway. The most significant one is that of parameter estimation of line constants and equivalent networks of external systems. Future developments may come up with standard techniques for the selection of sensor and telemetry system for on-line control. Furthermore, future advanced applications may require additional measurement equipment and more elaborate estimation and identification techniques. Notable among these are transient security monitoring, dynamic breaking, temperature estimation and others.

REFERENCES

1. F. C. Schweppe and J. Wildes, "Power System Static State Estimation, Part I: Exact Model," IEEE Trans. Power Apparatus and Systems, Vol. PAS-89, No. 1, pp. 120-125, January 1970.

2. F. C. Schweppe and D. B. Rom, "Power System Static State Estimation, Part II: Approximate Model," IEEE Trans. Power Apparatus and Systems. Vol. PAS-89, No. 1, pp. 125-130, January 1970.

3. F. C. Schweppe, "Power System Static State Estimation, Part III: Implementation," IEEE Trans. Power Apparatus and Systems, Vol. PAS-89, No. 1, pp. 130-135, January 1970.

4. R. E. Larson, W. F. Tinney, and J. Peschon, "State Estimation in Electric Power Systems, Part I: Theory and Feasibility," IEEE Trans. Power Apparatus and Systems, Vol. PAS-89, No. 3, pp. 345-352, March 1970.

5. R. E. Larson, W. F. Tinney, and L. P. Hajdu, "State Estimation in Electric Power Systems, Part II: Implementation and Applications," IEEE Trans. Power Apparatus and Systems, Vol. PAS-89, No. 3, pp. 353-363, March 1970.

6. A. S. Debs and R. E. Larson, "A Dynamic Estimator for Tracking the State of a Power System," IEEE Trans. Power Apparatus and Systems, Vol. PAS-89, No. 7, pp. 1670-1673, September/October 1970.

7. A. S. Debs and W. H. Litzenberger, "Experimental Evaluation of Tracking State Estimator," 1971 PICA Conference Proceedings, Vol. IEEE 71C26-PWR, pp. 348-355.

8. G. W. Stagg, J. F. Dopazo, O. A. Klitin, L. S. Van Slyck, "Techniques for the Real-Time Monitoring of Power System Operations," IEEE Trans. Power and Systems, Vol. PAS-89, No. 4, pp. 545-555, April 1970.

9. J. F. Dopazo, O. A. Klitin, G. W. Stagg, and L. S. Van Slyck, "State Calculation of Power Systems from Line-Flow Measurement," IEEE Trans. Power and Systems, Vol. PAS-89, No. 7, pp. 1698-1701, September/October 1970.

10. R. D. Masiello and F. C. Schweppe, "A Tracking Static State Estimator," IEEE Trans. Power Apparatus and Systems, Vol. PAS-90, No. 3, pp. 1205-1033, May/June 1971.

11. A. S. Debs, R. E. Larson, and L. P. Hajdu, "On-Line Sequential State Estimation for Power Systems," paper presented at the Power Systems Computation Conference in Grenoble, France, September 1972.

12. F. Arriatti and P. Ricci, "Methods for Electric Power System State Estimation: Comparative Analysis of Computing Simulations based upon Generalized Load Flow Tracking and Least Square Methods," PSCC Conference Proceedings, Grenoble, France, September 1972.

13. H. M. Merrill and F. C. Schweppe, "Bad Data Suppression in Power System Static State Estimation," IEEE PAS-90, pp. 2718-2725, November/December 1971.

14. J. F. Dopazo, O. A. Klitin, and A. M. Sasson, "State Estimation for Power Systems" Detection and Identification of Gross Measurement Errors," 8th PICA Conference Proceedings, pp. 313-318, June 1973.

15. A. S. Debs, "Power System Model Inaccuracies and Sequential Tracking Estimation," Technical Memorandum PAU949-3-1, prepared for Bonneville Power Administration by Systems Control, Inc., Contract 14-03-79968, Exhibit F, October 1971.

16. Richard Bauman, et. al., "The Main Factors of Influence in the Mathematical Modeling of Actual Conditions in Power Systems," ETZ-A, Vol. 87, No. 32, November 1966, pp. 818-823. Translated into English by approval of Bonnevill Power Administration, c/o H. W. Dommel.

17. F. C. Schweppe, "Role of System Identification in Electric Power Systems," paper presented at Power Systems Computation Conference held in Grenoble, France (September 1972).

18. A. S. Debs, "Estimation of Steady-State Power System Model Parameters," paper submitted for presentation at Winter Power Meeting, January 1974.

19. J. Bubenko and L. O. Petterson, "State Estimation in Electric Power Systems Using Filtering Techniques," paper presented at Power Systems Computation Conference held in Grenoble, France, September 1972.

20. T. Stuart and P. M. Anderson, "Sensitivity Analysis of State Estimation for Power Systems," paper presented at Winter Power Meeting, New York, January 1973.

21. K. A. Clements, et. al., "The Effects of Measurement Non-Simultaneity, Bias, and Parameter Uncertainty on Power System State Estimation," 1973 PICA Conference Proceedings, June 1973

22. R. K. Mehra, "Maximum Likeliehood Identification of Aircraft Parameters," presented at Joing Automatic Control Conference (JACC), Atlanta, Georgia, July 1970.

23. H. Cramer, Mathematical Methods of Statistics, Princeton University Press, 1963.

24. R. C. K. Lee, Optimal Estimation, Identification, and Control, MIT Press, Cambridge, Massachusetts, 1964.

A THEOREM ON CONSEQUENCES OF PLANNED DEMAND CHANGES IN A PARTIALLY FIXED-PRICE ENVIRONMENT

I. W. SANDBERG
Bell Laboratories
Murray Hill, New Jersey

ABSTRACT

We consider a multiproduct firm which, through price changes, plans to increase by specific amounts the demand for one or more goods, while holding constant the price of certain other goods as well as the demand for all remaining goods. It is shown under certain reasonable assumptions that when the price of a good is fixed, the demand for each other fixed-price good is either increased or unchanged and the consumer pays more for each good whose demand the firm plans to increase (relative to the case in which the price of that good is not fixed).

I. AN APPLICATION OF THE RESULT PROVED IN THE APPENDIX

Consider a firm which produces n products in which $n \geq 2$, and for each $i = 1,2,\ldots,n$ let p_i and q_i denote the price of product i and the external demand for product i, respectively.

With q the n-vector of demands and with p the n-vector of prices, suppose that

$$q = Q(p)$$

in which the demand function Q is assumed to possess the following properties.

(i) The demand for any good does not decrease if the price of that good remains constant and the prices of some other goods increase (i.e., all goods are weak gross substitutes).

(ii) If the price of one or more goods is changed, then there is at least one good for which either the price is decreased and the demand is increased or the price is increased and the demand is decreased.

Suppose that the firm is operating with price vector p^*, so that the demand vector is $Q(p^*)$ which we shall call q^*. Suppose also that it is desired to increase the demand for one or more goods[†], that the prices of certain other goods are fixed as a result of, for example, government regulation or fair trade agreements, and that the demands for all remaining goods are to be held constant.

More explicitly, suppose that the prices of goods $1, 2, \ldots, m$ are fixed and that, with $\ell \geq m$, it is desired to increase the demand for goods $(\ell+1)$, $(\ell+2), \ldots, n$ to $\hat{q}_{(\ell+1)}, \hat{q}_{(\ell+2)}, \ldots, \hat{q}_n$, respectively. The firm is free to change only the prices of goods $(m+1)$, $(m+2), \ldots, n$ and it is necessary to hold constant the demand for all nonfixed-price goods $(m+1)$, $(m+2), \ldots, \ell$. We shall consider the situation in which an increase in the prices of any incomplete set of the nonfixed-price goods causes the demand for some other nonfixed-price good to increase. Let us assume that this problem has a solution[††]. That is, given the n-vector $\hat{q}$ such that $\hat{q}_i > q_i^*$ for all $i \geq (\ell+1)$ and $\hat{q}_i = q_i^*$ for $i = (m+1), (m+2), \ldots, \ell$, we assume that there is a suitable (e.g., nonnegative) n-vector of prices p' such that $p'_i = p_i^*$ for $i = 1, 2, \ldots, m$ and $\hat{q}_i = Q_i(p')$ for all $i \geq (m+1)$.

Suppose now that the problem is changed only to the extent that m is replaced with $(m-1)$ and that the resulting problem also has a solution. Let us denote the solution n-vector by p''.

Here we wish to direct attention to two general relations concerning the demands for the fixed-price goods, the prices of the goods whose demand the firm desires to increase, and whether or not the price of some other good is fixed. According to an immediate application of the theorem proved in the Appendix we have $Q_i(p') \geq Q_i(p'')$ for all $i \leq (m-1)$ which indicates that the demand for any of the fixed-price goods $1, 2, \ldots, (m-1)$ is either increased or unchanged when the price of an additional good (i.e., the m^{th} good) is fixed. In addition, according to the theorem proved in the Appendix, we have $p'_i > p''_i$ for each good whose demand is to be increased.

[†] For example, the firm may be a public utility, and may wish to increase the demand for a new service in order to promote the growth of a system in anticipation of associated future reasonable profits and benefits to society, even if in the short run its profit in connection with the new service is negative. Very real situations of this type are not uncommon in, for example, the communications industry in which the value to each subscriber of a new communications service can increase very rapidly with increases in the total number of subscribers.

[††] It is not necessary to assume that the solution is unique.

That is, if the price of an additional good is fixed, the consumer will pay more for each good whose sale the firm plans to increase (relative to the case in which the price of the additional good is not fixed)[†].

It can be shown that $Q_i(p') \geq Q_i(p'')$ for all $i \leq (m-1)$ holds and we have the weaker inequalities $p_i' \geq p_i''$ for all $i \geq (\ell+1)$ if we drop the assumption that an increase in the prices of any incomplete set of the nonfixed-price goods causes the demand for some other nonfixed-price good to increase.

II. APPENDIX: THE THEOREM

The following notation is used throughout Section II. The symbol n denotes an integer such that $n \geq 2$. With r an arbitrary positive integer, E^r denotes the set of all real r-vectors, and for $v \in E^r$ and i an integer such that $1 \leq i \leq r$, v_i denotes the i^{th} component of v. For $v \in E^r$ and $w \in E^r$, the inequality $v \geq w$ means that $v_i \geq w_i$ for all i, and $v > w$ means that $v_i > w_i$ for all i.

The transpose of an arbitrary row vector $y = (y_1, y_2, \ldots, y_r)$ is denoted by y^{tr}. For any mapping $B(\cdot)$ of a subset E of E^r into E^r, let $B_i(x_1, x_2, \ldots, x_r)$ denote the i^{th} component of $B(x)$ for all $x = (x_1, x_2, \ldots, x_r)^{tr} \in E$ and all $i = 1, 2, \ldots, r$.

In the statement of the theorem we refer to A.1, the following combination of assumptions and definitions.

A.1: Let S denote a subset of E^n containing three elements p^*, p', and p'', and let S also contain all elements w of E^n such that for each i $w_i = p_i^*$, or $w_i = p_i'$, or $w_i = p_i''$. Let $Q(\cdot)$ denote a mapping of S into E^n such that

(i) for each i, $Q_i(v) \geq Q_i(w)$ whenever $v \in S$, $w \in S$, $v \geq w$ and $v_i = w_i$

(ii) for each $(v,w) \in S \times S$ such that $v \neq w$, there exists an index k such that $(v-w)_k\,[Q(v)-Q(w)]_k < 0$.

[†] This result is closely related to a theorem proved in Ref. 1 which generalizes a pioneering result[2] of Samuelson concerning a linear situation. Reference 1 deals in detail with the question of the existence of the three nonnegative vectors at issue, but (in the notation of the problem considered here) does not consider either the case in which more than one component of q is increased or the case in which Q can be discontinuous. Also, here assumption (ii) plays a key role. An alternative condition is used in Ref. 1 which is concerned primarily with input-output analysis. Conditions related to assumption (ii) are encountered in connection with equivalent definitions of certain classes of square matrices (see, for example, Ref. 3) and are used in the economics literature for other purposes as well (see, for example, Chapter 9 of Ref. 4).

Definition: With m any nonnegative integer such that $m < (n-1)$ we say that $Q(\cdot)$ of A.1 is (n-m)-coupled if and only if for each $(v,w) \in S \times S$ such that $v \neq w$, $v \leq w$, $v_i = w_i = p_i^*$ for all $i \leq m$, and $v_i = w_i$ for some $i \geq (m+1)$, we have $Q_\ell(w) > Q_\ell(v)$ for some $\ell \geq (m+1)$ such that $v_\ell = w_\ell$. If $m = (n-1)$, we say that $Q(\cdot)$ is 1-coupled.

Theorem: Let A.1 hold, and let q^* denote $Q(p^*)$. Let ℓ be a positive integer such that $\ell < n$, and let $\hat{q} \in E^n$ have the property that $\hat{q}_i = q_i^*$ for $i = 1,2,\ldots,\ell$ and $\hat{q}_i > q_i^*$ for $i = \ell+1, \ell+2,\ldots,n$.

Let m be a positive integer such that $m \leq \ell$. Suppose that the elements p' and p'' of S are such that $p_i' = p_i^*$ for $i \leq m$, $p_i'' = p_i^*$ for $i < m$, $\hat{q}_i = Q_i(p')$ for all $i \geq (m+1)$, and $\hat{q}_i = Q_i(p'')$ for all $i \geq m$. Suppose also that $Q(\cdot)$ is (n-m+1)-coupled. Then $p_i'' < p_i^*$ for all $i \geq m$, and $p_i'' < p_i'$, for all $i \geq (m+1)$ provided that $Q(\cdot)$ is (n-m)-coupled.

Proof of the Theorem

We shall use the following lemma which is a special case of a result proved in Ref. 5.

Lemma: Let a and b be elements of E^r such that $a \neq b$, and let S' denote the set $\{v \in E^r:$ for each i $v_i = a_i$ or $v_i = b_i\}$. Let $A(\cdot)$ map S' into E^r such that

(i) $A(a) \geq A(b)$

(ii) for each i, $A_i(b) \geq A_i(v)$ for all $v \in S'$ such that $b \geq v$ and $b_i = v_i$

(iii) if $r > 1$, either for each $v \in S'$ such that $v \neq b$, $v \leq b$, and $v \nless b$, we have $A_\ell(b) > A_\ell(v)$ for some index ℓ such that $b_\ell = v_\ell$, or for each $v \in S'$ such that $v \neq a$, $v \geq a$, and $v \ngtr a$, we have $A_\ell(v) > A_\ell(a)$ for some index ℓ such that $a_\ell = v_\ell$.

(iv) for each $v \in S'$, except $v = a$, there exists an index k such that

$$(v-a)_k[A(v)-A(a)]_k < 0.$$

Then $a < b$.[†]

[†] This lemma is closely related to earlier results which assert that if $Q(\cdot)$ is defined on a rectangular region and other conditions are met, then $a \leq b$ (see Refs. 6 and 7). For our purposes it is essential to be able to prove that $a < b$. In using the lemma stated above, rather than building on an existing result, we arrive at a theorem in which the domain of $Q(\cdot)$ can be restricted to a finite set of n-vectors, in which each element of each n-vector can be interpreted as a price which is an element of at least one of three special price vectors of interest.

Since

$$Q_i(p_1^*,\dots,p_{m-1}^*,p_m'',\dots,p_n'') - Q_i(p_1^*,p_2^*,\dots,p_n^*) = \hat{q}_i - q_i^*$$

for all $i \geq m$, using the lemma and the hypothesis that $Q(\cdot)$ is $(n-m+1)$-coupled, we have $p_i'' < p_i^*$ for all $i \geq m$. In particular,

$$Q_i(p_1^*,\dots,p_{m-1}^*,p_m^*,p_{m+1}'',\dots,p_n'') - Q_i(p_1^*,\dots,p_{m-1}^*,p_m'',\dots,p_n'') \geq 0 \qquad (1)$$

for all $i \geq (m+1)$, and, because $Q(\cdot)$ is $(n-m+1)$-coupled, (1) is satisfied with inequality for some $i \geq (m+1)$. The left side of (1) is equal to

$$Q_i(p_1^*,\dots,p_{m-1}^*,p_m^*,p_{m+1}'',\dots,p_n'') - Q_i(p_1^*,\dots,p_m^*,p_{m+1}',\dots,p_n')$$

for all $i \geq (m+1)$. Therefore, using the lemma and the hypothesis that $Q(\circ)$ is $(n-m)$-coupled, we have $p_i'' < p_i'$ for all $i \geq (m+1)$, which completes the proof.

REFERENCES

1. Sandberg, I. W., "A Global Nonlinear Extension of the Lechatelier-Samuelson Principle for Linear Leontief Models", to be published in the Journal of Economic Theory.
2. Samuelson, P. A., The Collected Scientific Papers of Paul A. Samuelson, J. E. Stiglitz, Ed., M.I.T. Press, Cambridge, Mass. 1965, Vol. 1, Chapter 42.
3. Fiedler, M., and Ptak, V., "On Matrices with Non-Positive Off-Diagonal Elements and Positive Principal Minors", Czech. Math. J., 12, No. 3 (1962), pp. 382-400.
4. Arrow, K. J. and Hahn, F. H., General Competitive Analysis, Holden-Day, Inc., San Francisco, 1971.
5. Sandberg, I. W., "On the Mathematical Theory of Interaction in Social Groups", to be published.
6. Gale, D. and Nikaido, H., "The Jacobian Matrix and Global Univalence of Mappings", Math. Ann. 159, No. 2 (1965).
7. Rheinboldt, W. C., "On Classes of n-Dimensional Nonlinear Mappings Generalizing Several Types of Matrices", pp. 501-546 of Numerical Solution of Partial Differential Equations - II, B. Hubbard, Ed., Academic Press, Inc., New York and London, 1971.

PRODUCTION CONFIGURATIONS, TECHNOLOGY SETS, AND DISCRETE SYSTEMS

Jose Antonio Ortega
COPPE, Universidade Federal
do Rio de Janeiro, Brasil

R. Jeffrey Leake
Dept. of Electrical Engineering
University of Notre Dame

ABSTRACT

The purpose of this talk is to report on recent research relating economic growth theory and the maximum principle of discrete optimal control. A number of authors have derived conditions relating equilibrium price conditions and the maximum principle using various mathematical models which are uncommon in the control literature. The authors' work has involved a formal study of various models used and a subsequent determination of and solution of what appears to be an essential problem within the usual discrete time system framework.

The most commonly used model in economic growth theory is a Technology set $T \subset E^n \times E^n$ in which feasible paths $\{x_i\}$ satisfy $(x_i, x_{i+1}) \in T$.

Another model which introduces an extra degree of freedom is a Production Configuration $P \subset E^n \times E^m \times E^n$ in which feasible paths $\{x_i\}$ satisfy $(x_i, v_i, x_{i+1}) \in P$ for some sequence $\{v_i\}$.

An associated discrete time system model S is $x_{i+1} = f(x_i, u_i)$ where $(x_i, u_i) \in X \subset E^n \times E^m$.

It can be shown that each model of one type uniquely determines a model of the other type. Given S for example, P may be taken as the graph of f; and given P we may determine an S by making $X = \{ (x,u) : u = (v,y), (x,v,y) \in P\}$ and defining $x_{i+1} = y_i$. Thus, systems S yield restricted production configurations which are the graph of a function and production configurations can always be reduced to trivial systems with complicated control constraint sets.

Convexity assumptions are commonly made for T and P models and in a specific sense this corresponds to linear affine systems S.

It has been shown that a generalization of the usual discrete maximum principle can be obtained for systems S if X is assumed convex, and this corresponds to a solution of an optimal control problem in which the control constraint set depends on the system state. Using this result, stronger equilibrium price conditions are obtained for economic growth problems.

REFERENCES

J. A. Ortega, "Um Principio Do Maximo Discreto Para Modelos Economicos de Producao." D.Sc. Thesis, COPPE, Universidade Federal do Rio de Janeiro (1973).

J. A. Ortega and R. J. Leake, "Discrete Maximum Principle with State Constrained Control," (1973). Submitted for publication.

A DYNAMIC SALES MODEL FOR REUSABLE COMMODITIES

JOSEPH G. SCHATZ
Bell Laboratories, Whippany, New Jersey

1. INTRODUCTION

The sales of a commodity are classically estimated at a given instant of time by use of a Demand Curve, which portrays sales volume as a function of unit price. These curves, constructed through an analysis of both consumer income and consumer product preference, are in reality functions of both past sales* and many independent parameters. Typical parameters may include past and present advertising expenditures, interest rates, etc. Thus, the sales of a product over a period of time is an example of a dynamic process. The Demand Curve, then, represents a snapshot of this process (as a function of price) at a specific time and for predetermined values of the parameters.

In many instances an estimate of the sales volume of a commodity at future instants of time is required for production planning. In order to obtain such an estimate, a model of the dynamic sales process must be formulated. This formulation may either require the construction of a large number of Demand Curves (one curve for each instant of time and combination of parameter and sales history values) or the derivation of a Demand Function, which is an analytic expression for demand as a function of both past sales and the pertinent parameters. The former approach, although conceptually straightforward, has the drawback of being tedious to develop and cumbersome to implement on a digital computer. Alternately, the latter approach may not be analytically feasible but is potentially easier to implement on a computer.

The purpose of this paper is to develop a mathematical model for the dynamic sales process of a commodity which is used on a temporary basis and, therefore, can be subsequently reused. Thus, the sales for such a commodity over a period of time are a function not only of past sales but also of the reusability and durability properties of the product. The modeling technique, which is described, makes use of combination of Demand Curve and Demand Function approaches and is thus referred to as the Hybrid approach.

---

* For many products, this functional dependence may be neglected.

2. DYNAMIC SALES MODEL

This Hybrid model utilizes a set of Demand Curves which are constructed under the assumption that the commodity is not reusable. One curve, is required for each forecast instant. An analytic expression is then developed which modifies these curves in such a manner as to include the effects of reusability, retirements and ultimately of past sales. Thus the Hybrid model has the advantage of both dramatically reducing the number of Demand Curves required for simulation and being easy to implement on a digital computer. The Hybrid model is composed of a set of discrete-time, non-linear equations which have the following form:

$$x_i = \{D_i - f(e_i, a_i, S_{i-1})\}u_o(D_i - f) ; \qquad (1)$$

$$i = 0,1,2,\ldots, ; \quad S_{-1} = 0$$

$$S_i = -\{D_i - f(e_i, a_i, S_{i-1})\}u_o(f - D_i) ; \qquad (2)$$

where

x_i = the number of units sold at time i

D_i = the number of units demanded at time i if no units are available from either reuse or surplus stock. This quantity is assumed to be given.

e_i = the net number of units which may be reused at time i; i.e., the number of reusable units minus the number of retired units

a_i = the number of units which were placed in an application at a time previous to i and are still being used in that application

S_i = the number of units which will be idle or surplus at time i

$u_o(\cdot)$ = unit step function

$f(\cdot)$ = scalar-valued non-linear function.

At any instant of time the quantities e_i and a_i are functions of the number of units which have been previously purchased and are presently being reused or retired. Thus, these two quantities are generated through use of separate retirement and reuse models. These models utilize previous outputs of the dynamic sales model as inputs. In this paper both the retirement and reuse models are assumed to be given. It should be noted that the consistency constraints on the number of units which can be reused or retired are assumed to be embodied in the reuse and retirement models respectively. A flow diagram showing the relationship between the dynamic sales model, the reuse model and the retirement model is portrayed in Figure 1.

3. CONCLUSION

In summary, the Hybrid model represents an analytically and computationally tractable approach to simulating the dynamic sales process of a reusable commodity. This approach dramatically reduces the number of Demand Curves required for simulation and avoids development of complex demand functions. Furthermore, the form of the model lends itself to the development of optimal production schedule strategies for products of this type.

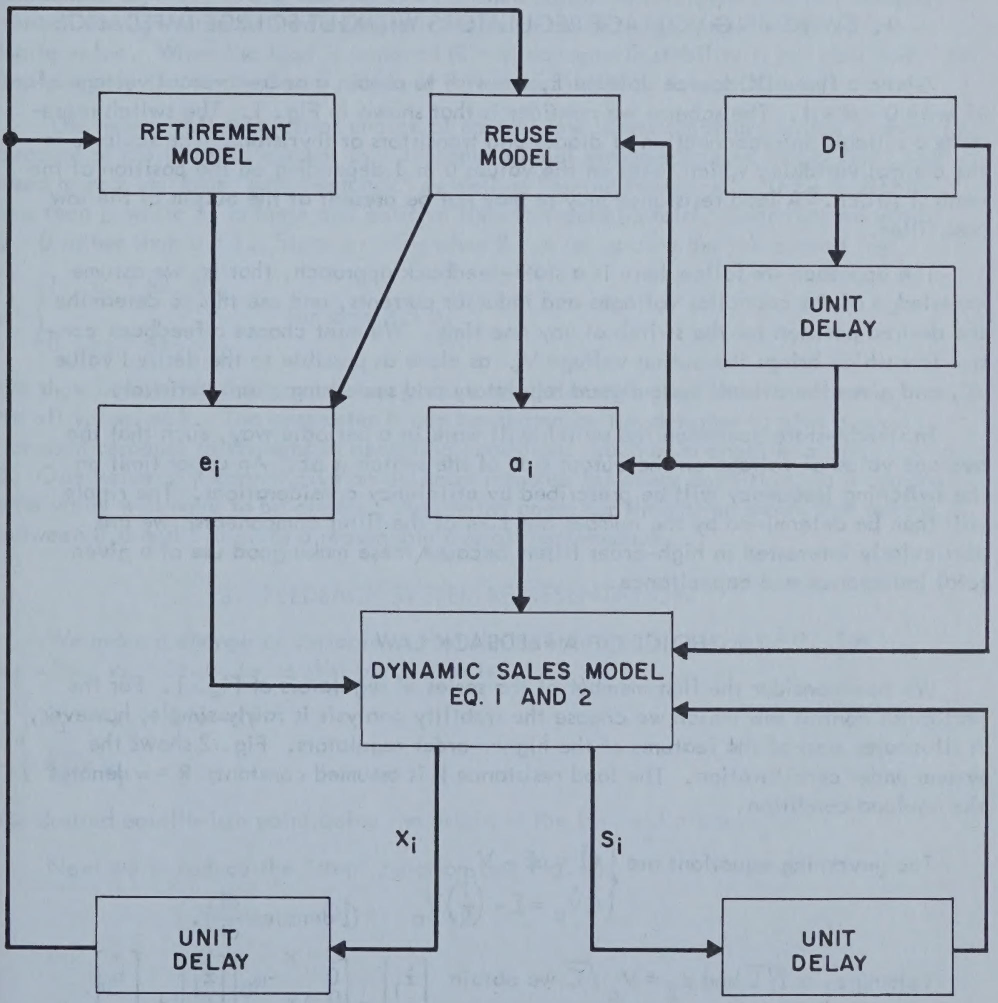

RETIREMENT MODEL
REUSE MODEL
D_i
UNIT DELAY
e_i
a_i
DYNAMIC SALES MODEL EQ. 1 AND 2
X_i
S_i
UNIT DELAY
UNIT DELAY

FIGURE I

STABILITY CONSIDERATIONS IN SWITCHING VOLTAGE REGULATORS

DR. JONATHAN R. WOOD
Central Research Laboratories
Texas Instruments Incorporated
Dallas, Texas

A class of switching voltage regulators is considered from a stability viewpoint. High-order filtering networks make good use of a given total inductance and capacitance, but for these networks phase-plane methods cannot be used for analyzing stability. We show how the Positive Operator Theorem can be used, both directly and indirectly by providing Lyapunov functions.

1. SWITCHING VOLTAGE REGULATORS WITHOUT SOURCE IMPEDANCE

Given a fixed DC source voltage E, we wish to obtain a near-constant voltage of αE with $0 < \alpha < 1$. The scheme we consider is that shown in Fig. 1. The switch represents a suitable interconnection of diodes and transistors or thyristors. The scalar u is the control variable, which takes on the values 0 or 1 depending on the position of the control switch. A load resistance may or may not be present at the output of the low pass filter.

The approach we follow here is a state-feedback approach, that is, we assume knowledge of the capacitor voltages and inductor currents, and use this to determine the desired position for the switch at any one time. We must choose a feedback control law which brings the output voltage V_o as close as possible to the desired value αE, and gives the overall system good regulatory and smoothing characteristics.

In steady-state operation the switch will work in a periodic way, such that the average value of voltage on the output side of the switch is αE. An upper limit on the switching frequency will be prescribed by efficiency considerations. The ripple will then be determined by the number and size of the filter components: we are particularly interested in high-order filters because these make good use of a given total inductance and capacitance.

2. CHOICE OF A FEEDBACK LAW

We now consider the first member of the series of regulators of Fig. 1. For the particular control law which we choose the stability analysis is fairly simple; however, it illustrates most of the features of the higher-order regulators. Fig. 2 shows the system under consideration. The load resistance R is assumed constant; $R = \infty$ denotes the no-load condition.

The governing equations are

$$\begin{cases} L\dot{I} = uE - V_o \\ C\dot{V}_o = I - \left(\frac{1}{R}\right)V_o \end{cases} \qquad (\dot{I} \text{ denotes } \frac{dI}{dt}).$$

Letting $z_1 = I\sqrt{L}$ and $z_2 = V_o\sqrt{C}$ we obtain

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & -\omega_0 \\ \omega_o & -\omega_1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} b \\ 0 \end{bmatrix} u$$

i.e.

$$\dot{\underline{z}} = \underline{\underline{A}}\,\underline{z} + \underline{b}u$$

where $\omega_0 = \left(\frac{1}{\sqrt{LC}}\right)$, $\omega_1 = \left(\frac{1}{RC}\right)$, $b = \left(\frac{E}{\sqrt{L}}\right)$. By introducing time- and amplitude-scaling we may assume without loss of generality that $E = L = C = 1$. (Let $\tau = \omega_0 t$ and $y_i = \frac{\omega_0 z_i}{b}$).

We want to make u a function of the state vector $\underline{z}$, (i.e. the control u a function of the measured state variables z_1 and z_2) so that z_2 comes as close as possible to its desired value α (where $0 < \alpha < 1$). That is, we want $\lim z_2(t) = \alpha$, as $t \to \infty$, for any initial condition $(z_1(0), z_2(0))$. We call this global asymptotic stability about $z_2 = \alpha$. Now if $z_2 < \alpha$ one would expect that the switch should be in the $u = 1$ position, and if $z_2 > \alpha$ that the switch should be in the $u = 0$ position. This leads us to try the feedback control $u = 1$, $z_2 < \alpha$; $u = 0$, $z_2 > \alpha$. (For brevity we avoid discussion of what happens at $z_2 = 0$.) Using the methods outlined below we can show that this feedback control does give the desired asymptotic stability about $z_2 = \alpha$ provided that R has a finite value. When the load is removed ($R = \infty$) asymptotic stability is not obtained. We analyze only the no-load case, since it is the most difficult to stabilize.

The reason that this initial choice of control law is unsatisfactory is that it takes no account of the inductor current z_1. In conventional servomechanism theory terms, we need to provide some "rate feedback" as well as "output feedback". If z_2 is slightly less than α while z_1 is large and positive then it should be fairly clear that we want $u = 0$ rather than $u = 1$. Since $z_1 = \dot{z}_2$ when $R = \infty$ let us consider the control law

$$u = \begin{cases} 1, & z_2 + \beta\dot{z}_2 < \alpha \\ 0, & z_2 + \beta\dot{z}_2 > \alpha \end{cases} \qquad \text{for some } \beta > 0.$$

We show below that this control law gives a regulator with good stability about $z_2 = \alpha$ for all values of R. The parameter β can be chosen by the designer to give a good transient response to changes in operating conditions, such as changes in α, in E, or in R. One value of β cannot give an optimum response for every condition, so a compromise value will have to be chosen. An analog computer simulation shows that setting β between 0.8 and 1.0 gives a reasonable overall performance.

3. FEEDBACK SYSTEM REPRESENTATION

We make a change of variable so that the desired stability is about 0. Let $x_1 = 2z_1$, $x_2 = 2z_2 - 2\alpha$ so that our system is

$$\begin{cases} \dot{x}_1 = -x_2 + 2(u - \alpha) \\ \dot{x}_2 = x_1 \end{cases}$$

the desired equilibrium point being the origin in the (x_1, x_2) plane.

Next we introduce the "step" function (see Fig. 3):

$$\underset{(-a,b)}{\text{stp}}\ x = \begin{cases} -a, & x < 0 \\ 0, & x = 0 \\ b, & x > 0 \end{cases}$$

Then our system is

$$\begin{cases} \dot{x}_1 = -x_2 - \underset{(-2+2\alpha,\ 2\alpha)}{\text{stp}} (\beta\dot{x}_2 + x_2) \\ \dot{x}_2 = x_1 \end{cases}$$

We assume henceforth that the value of α is known and abbreviate $\underset{(-2+2\alpha,2\alpha)}{\mathrm{stp}}$ to stp.

Some remarks are necessary concerning the use which we make of the stp function. This is a discontinuous function, and a statement such as

$$\frac{d}{dx}\,|x| = \underset{(-1,1)}{\mathrm{stp}}\;x = \mathrm{sgn}\,x$$

requires rigorous mathematical treatment in order to make strict sense for all values of x. However, we shall use the stp function merely for its notational and conceptual convenience, since it is the limit of various classes of continuous functions such as sat kx or tanh kx, where k is an arbitrarily large positive real number; such functions are more accurate descriptions of a "real-world" switching function than stp x is. In order to put our treatise on a firm mathematical footing we need only replace stp x by tanh kx (or some other suitably smooth function) each time it is used, and show that the relevant conclusions still hold true. In fact, many of our results hold true if stp x is replaced by f (x) where f is any monotone nonlinearity with f (0) = 0. (A monotone function f(0) is one for which $f(x_1) \leq f(x_2)$ whenever $x_1 \leq x_2$.)

Now by writing our system in the form

$$\dot{x} + x = -\,\mathrm{stp}\,(\beta x + x)$$

and taking Laplace transforms, we can represent this in the usual feedback system form of Fig. 4. Consider now a feedback system with transfer function G(s) in the forward path, and nonlinearity $f(\sigma)$ in the feedback path, as in Fig. 5. Assume that f(0) = 0. We can associate with this the linear system of Fig. 6, with feedback gain k.

Aizerman's Conjecture states that for the system of Fig. 5 we would expect to have global asymptotic stability if the associated system of Fig. 6 is globally asymptotically stable for all values of k lying in the range of values taken on by $\frac{f(\sigma)}{\sigma}$ as σ varies along the real line. This method of investigating the stability of a feedback system is called the method of total gain linearization (see Ref. 4). Aizerman's Conjecture is not true in general, though it appears to be true for the type of systems considered here, and it can be shown to be true for second-order systems. The nonlinearity stp σ lies in the first and third quadrants, with $\frac{\mathrm{stp}\,\sigma}{\sigma}$ taking on all values between 0 and $+\infty$. The total gain linearization stability conditions are thus fulfilled if the Nyquist locus of G(s) does not cross the negative real axis. This is so for the system of Fig. 4 for which the Nyquist locus is as shown in Fig. 7. The Circle Criterion is not useful here, because the interior of the disc is the left half plane, and the Nyquist locus enters this region.

4. STABILITY BY THE POSITIVE OPERATOR THEOREM

A positive operator is an operator with input u(t) and output y(t), where $0 \leq t < \infty$, for which $\int_0^T uy\,dt \geq 0$ for all $T \geq 0$, and for which y(t) = 0 for all t whenever u(t) = 0 for all t. This last requirement can be written G 0 = 0. If the operator is a convolution operator, and if it can be represented by a rational transfer function G(s), then it can be shown that positivity is equivalent to the requirement that G(s) have no poles in the right half plane, that any poles on the imaginary axis be simple with real positive residues, and that Re $G(j\omega) \geq 0$ for all ω, i.e. the Nyquist locus of G(s) must lie entirely in the right half plane. An equivalent requirement is that Re $G(j\omega) \geq 0$ and if

$G(s) = \frac{q(s)}{p(s)}$ then $p(s) + q(s)$ must be strictly Hurwitz, (i.e. all its zeros must lie in $\text{Re } s < 0$). Such functions are called positive real and play an important role in electrical network synthesis, since the driving point impedance of a linear passive network is positive real, and any positive real function is the driving point impedance of an RLC network.

The Positive Operator Theorem gives a sufficient condition for input-output stability of the feedback system of Fig. 8 in which $u_1, y_1, u_2, y_2, v_1, v_2$ are all functions of t for $0 \le t < \infty$. The functions v_1 and v_2 are the inputs, and can be used to represent driving functions, driving noise, or initial condition responses. By choosing v_1 and v_2 appropriately we can use the Positive Operator Theorem to obtain conclusions about the behavior of u_1, y_1, u_2, y_2 as $t \to \infty$. To do this we need a measure of the "size" of a function of time $x(t)$; we use the L_2-norm

$$||x(t)|| = \left(\int_0^\infty x^2 \quad dt \right)^{\frac{1}{2}} .$$

Using this norm we can only handle functions $x(t)$ for which $\int_0^\infty x^2 \ dt$ is finite; the set of all such functions is denoted L_2. There are many functions of interest which are not in L_2, for example the constant functions. This difficulty is overcome by using truncated functions; that is, functions which are zero after some time T. It can be shown that for any function $x(t) \in L_2$ for which $\dot{x}(t)$ is bounded or square-integrable (i.e. $\int_0^\infty \dot{x}^2 \ dt < \infty$, we must have $\lim_{t \to \infty} x(t) = 0$. An operator G is said to be bounded if there exists a real number M such that $\frac{||Gx||}{||x||} < M$ for all x. An operator G with input u(t) and output y(t) is said to be strictly positive if $G - \eta I$ is positive for some $\eta > 0$, I being the identity operator.

The Positive Operator Theorem (Refs. 8, 4, 7) states that if G_1 and G_2 are positive with one of them being strictly positive and bounded, then u_1, y_1, u_2, y_2 are all in L_2 whenever v_1 and v_2 are in L_2, and there exist positive constants ρ_1 and ρ_2 such that

$$||u_1||, ||y_1||, ||u_2||, ||y_2|| \le \rho_1 ||v_1|| + \rho_2 ||v_2|| .$$

Now consider application of the Positive Operator Theorem to the feedback system of Fig. 4, which represents our second-order lossless voltage regulator. $G_1(s) = \frac{\beta s + 1}{s^2 + 1}$ is not a positive operator, since its Nyquist locus enters the left half plane, as depicted in Fig. 7. However, $G_2(\sigma) = \text{stp } \sigma$ is "very strongly positive", and by introducing two multiplying factors into the loop we can make use of this fact, so that we end up with two positive operators. Fig. 9 depicts the same system. We have introduced a multiplier $(\gamma s + 1)$ in the forward path, and $(\gamma s + 1)^{-1}$ in the feedback path, where $\gamma \beta = 1$. Introducing these factors has no effect provided that the initial conditions of $(\gamma s + 1)^{-1}$ are correctly chosen: there will be a linear relationship between the initial conditions of $(\gamma s + 1)^{-1}$ and those of $H_1(s) = \frac{(\beta s + 1)(\gamma s + 1)}{s^2 + 1} = 1 + \frac{(\beta + \gamma)s}{s^2 + 1}$. Now for $H_1(s)$ we have $\text{Re } H_1(j\omega) = 1 > 0$, so $H_1(s)$ defines a positive operator. To show that the feedback operator H_2 is positive, consider the operator shown in Fig. 10, where $f(\sigma)$ is any first- and third-quadrant nonlinearity, $\gamma > 0$, and $w(0) = 0$. Let $\int_0^x f(\sigma) \, d\sigma = F(x)$, ≥ 0 for all x. Then

$$\int_0^T uy \ dt = \int_0^T (w + \gamma \dot{w}) \ f(w) \ dt = \int_0^T w \ f(w) \ dt + \gamma \int_0^{w(T)} f(w) \ dw$$

$$= \int_0^T w \ f(w) \ dt + \gamma F(w(T)) - \gamma F(w(0)) \quad \ge 0.$$

From the Positive Operator Theorem we now conclude (see Ref. 10 for details) that our second-order voltage regulator is stable for any $\beta > 0$.

We make two comments here about the multiplier $(1+\gamma s)$, $\gamma\beta = 1$. Firstly, it is easy to show that in this case this is the only such multiplier which will make $\frac{\beta s+1}{s^2+1}$ become positive real. Secondly, we could have predicted that such a multiplier exists by making use of a theorem of Brockett and J. L. Willems (Ref. 1) which says that the feedback system of Fig. 6 is asymptotically stable for any feedback gain k satisfying $0 < k < k^*$, if and only if there exists a positive real function Z(s) such that $Z(s)\left[G(s) + \frac{1}{k^*}\right]$ is positive real.

5. HIGHER-ORDER REGULATORS

Fig. 11 shows the fourth-order regulator. In steady-state operation we want $z_2 = z_4 = \alpha$ and $\dot{z}_2 = \dot{z}_4 = 0$. The natural extension of the control law we chose before is

$$u = \begin{cases} 1\ , & \text{if } \beta_1\dot{z}_2 + \beta_2(z_2-\alpha) + \beta_3\dot{z}_4 + \beta_4(z_4-\alpha) < 0 \\ 0\ , & \text{otherwise,} \end{cases}$$

where β_1, β_2, β_3, β_4 are suitably chosen positive constants. This leads us to the feedback representation of Fig. 12, which is of the form of Fig. 5. It is easy to show that the associated feedback system of Fig. 6 is stable for all $0 < k < \infty$ if and only if $\beta_3 \le 2\beta_1$, $\beta_3{}^2 < \beta_1(\beta_1+\beta_3)$, and $\beta_1\beta_4 \le \beta_2\beta_3$. In particular we note that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ satisfies these requirements. To establish stability by the Positive Operator Theorem for this case, we need to find a multiplier Z(s) such that G(s)Z(s) is positive real, where

$$G(s) = \frac{s^3+s^2+2s+2}{s^4+3s^2+1} = \frac{(s+1)\ (s^2+2)}{(s^2+2.61)\ (s^2+0.38)}\ .$$

The simplest such multiplier Z(s) is of the form $\frac{s^3+as^2+bs+c}{s^2+ds+e}$, and positive realness of G(s) Z(s) leads to requirements on a,b,c,d,e. To ensure that $Z^{-1}(s)$ followed by stp x yields a positive operator we make use of a theorem of O'Shea (Refs. 2,3) which says that the operator shown in Fig. 13 with input u(t) and output y(t) and Z(s) rational, is a positive operator for any monotone nonlinearity $f(\sigma)$ for which $f(0) = 0$, if and only if $Z(s) = g_o + \gamma s - \hat{g}(s)$ where $\gamma \ge 0$, $\hat{g}(s) = \int_0^\infty g(t)e^{-st}dt$, $g(t) \ge 0$, and $g_o \ge \hat{g}(0) = \int_0^\infty g(t)\,dt$. Applying this theorem to our Z(s) we obtain another set of constraints on a, b,c,d,e. After some trial and error one can find a set of values for a,b,c,d,e which satisfy both sets of constraints. One such set yields

$$Z(s) = \frac{s^3+200s^2+96s+4}{s^2+1.45s+0.5}\ .$$

Using this we conclude that the fourth-order regulator is globally asymptotically stable when $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. The practical implications of this are that it is just as easy to stabilize a fourth-order regulator as it is to stabilize a second-order regulator, and an analog computer simulation readily demonstrates this.

6. THE DISSIPATIVE SYSTEMS APPROACH

We now show that the second-order regulator is stable using concepts from the theory of Dissipative Systems (Ref. 5), which is closely allied with the theory of

Positive Operators. Let G be an operator (or system) with input u(t) and output y(t), $0 \leq t < \infty$, defined by the equations $\dot{\underline{x}}(t) = \underline{f}(\underline{x},u)$, $y = g(\underline{x})$, and $\underline{x}(0) = \underline{x}_0$. The vector $\underline{x}(t)$ is called the state vector for G. Let $w(u,y)$ be a real-valued function of u and y. Then G is said to be dissipative with respect to the supply-rate w(u,y) if there exists a nonnegative function $V(\underline{x})$ with $V(\underline{0}) = 0$ such that $\dot{V} - w \leq 0$. V is called the storage function for G. The only supply rate w which we consider is the product uy, so that when we refer to an operator as dissipative we have this supply rate in mind. In this case one can show that an operator is dissipative if and only if it is a positive operator.

Consider for instance a one-port RLC electrical network with input current u(t) and voltage y(t). The operator G mapping u(t) into y(t) represents the impedance of the network; expressed as a transfer function G(s) it is a positive real function. The requirement for G to be a positive operator, i.e., $\int_0^T uy\,dt \geq 0$, says that the net energy absorbed by the network is at all times positive, the concurrent requirement being that G maps zero into zero, i.e. that no initial energy is stored in the network. If G is viewed as a dissipative operator the storage function V is the energy stored in the network, and the supply rate w = uy is the rate at which energy is supplied. The requirement $\dot{V} - uy \leq 0$ says that the rate of increase of stored energy is not greater than the rate of supply, (because of dissipation within the network). This requirement is independent of the initial conditions within the network.

Now consider the feedback connection of Fig. 8 with $v_1 \equiv v_2 \equiv 0$. Let G_1 be a dissipative operator with state vector $\underline{x}_1$ and storage function $V_1(\underline{x}_1)$, and let G_2 be dissipative with $\underline{x}_2$ and $V_2(\underline{x}_2)$. Then $\dot{V}_1 - u_1y_1 \leq 0$, and $\dot{V}_2 - u_2y_2 \leq 0$. But $u_1 = -y_2$ and $u_2 = y_1$, so if we let $V(\underline{x}_1, \underline{x}_2) = V_1(\underline{x}_1) + V_2(\underline{x}_2)$ we obtain $\dot{V} = \dot{V}_1 + \dot{V}_2 = \dot{V}_1 - u_1y_1 + \dot{V}_2 - u_2y_2 \leq 0$. Thus, a Positive Operator Theorem proof of stability provides us with a Lyapunov function for use with the Lyapunov Theorem, provided that the positive operators can be described in dissipative operator form.

A state-space description of the second order regulator of Fig. 4 with $\beta = 1$ is

$$\left\{\begin{array}{lll} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} & = & \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u \\ z & = & \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad u = -\mathrm{stp}\, z \end{array}\right.$$

i.e. $\dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{b}\,u, \quad z = \underline{g}\,\underline{x}, \quad u = -f(z)$

where $\underline{g}(\underline{I}s - \underline{A})^{-1}\underline{b} = \dfrac{s+1}{s^2+1}$.

A state-space description of the system of Fig. 9 is a non-minimal realization of the system of Fig. 4. We have

$$\left\{\begin{array}{lll} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} & = & \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u \\ y & = & \begin{bmatrix} 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + u \\ \dot{z} & = & -z + y \text{ with } z(0) = x_1(0) + x_2(0) \qquad u = -\mathrm{stp}\, z \end{array}\right.$$

i.e. $\begin{cases} \dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{b}\,u, & y = \underline{c}\,\underline{x} + d\,u \\ \dot{z} = -z + y, & u = -f(z) \end{cases}$

where $\underline{c}(Is - \underline{a})^{-1}\,\underline{b} + d = 1 + \frac{2s}{s^2+1}$.

Note that

$$z(t) \equiv x_1(t) + x_2(t),$$

since

$$\dot{z} + z = y = 2x_1 + u = (\dot{x}_1 + \dot{x}_2) + (x_1 + x_2) .$$

Now it can be shown (see Ref. 5) that a storage function $V_1(x)$ for a linear system such as $(1 + \frac{2s}{s^2+1})$ is $\frac{1}{2}\underline{x}'\underline{K}\,\underline{x}$ where $\underline{K} = \underline{K}'$ satisfies the matrix equation

$$\underline{K}\,\underline{A} + \underline{A}'\underline{K} + (\underline{K}\,\underline{b} - \underline{c}')\,(2d)^{-1}\,(\underline{b}'\underline{K} - \underline{c}) = \underline{0}.$$

In this case this equation has a unique solution $\underline{K} = 2\underline{I}$, so that the storage function for $(1 + \frac{2s}{s^2+1})$ is $V_1(\underline{x}) = \underline{x}'\underline{x} = x_1^2 + x_2^2$. Furthermore (see Refs. 5,7) the unique storage function for the feedback operator depicted in Fig. 14 is $V_2(z) = z \text{ stp } z$. Since $z = x_1 + x_2$, we have $V_2(\underline{x}) = (x_1 + x_2) \text{ stp } (x_1 + x_2)$. Thus a Lyapunov function for the second-order regulator described by

$$\{ \; \dot{x}_1 = -x_2 - \text{stp}(x_1 + x_2), \quad \dot{x}_2 = x_1$$

is $\quad V(\underline{x}) = V_1(\underline{x}) + V_2(\underline{x}) = x_1^2 + x_2^2 + (x_1 + x_2) \text{ stp } (x_1 + x_2)$.

Differentiating we get

$$\begin{aligned} \dot{V}(\underline{x}) &= 2x_1\dot{x}_1 + 2x_2\dot{x}_2 + (x_1 + x_2) \text{ stp } (x_1 + x_2) \\ &= -\text{stp}^2 (x_1 + x_2) - (x_1 + x_2) \text{ stp } (x_1 + x_2) \\ &\le 0, \; = 0 \text{ if and only if } x_1 + x_2 = 0. \end{aligned}$$

But $\quad \dot{x}_1 + \dot{x}_2 = x_1 - x_2 - \text{stp}(x_1 + x_2)$

$$= 0 \text{ on } \{x_1 + x_2 = 0\} \quad \text{only at } x_1 = x_2 = 0 ,$$

i.e. $\dot{V}(\underline{x})$ is identically zero only along the solution $\underline{x} = \underline{0}$. Thus, by a Lyapunov Theorem (see Refs. 6,7) we obtain the desired global asymptotic stability about $\underline{x} = \underline{0}$. Notice that for Fig. 9 we have proved stability for any initial conditions in the forward path and any initial conditions in the feedback path; this includes the case where there is a linear relationship between these two sets of initial conditions, this case being the system of Fig. 4.

Reference 10 shows how this Dissipative Systems approach can be extended to deal with the case of a regulator whose voltage source has a finite series resistance, as depicted in Fig. 15. The evolution equations in this case take the form

$$\dot{\underline{z}} = (\underline{A}o + u\underline{A}_1)\;\underline{z} + \underline{b}u .$$

7. IMPLICATIONS FOR DESIGN

We now discuss briefly some of the practical aspects of implementing the control laws we have been considering. Fig. 16 shows a schematic for one possible implementation of the type of regulator we have been discussing. The two-position control switch is effected by means of the return-path diodes D_1 and D_2, the power transistor T_1, and the driver transistor T_2. The two resistors marked R_1 are used to obtain a measure of half of the output voltage V_0. This and half the desired output voltage of αE are fed into the operational amplifier network which gives $k(\beta s+1)(V_0-\alpha E)$, i.e. $k(\beta\dot{V}_0+V_0-\alpha E)$, where $\beta = R_2C_2$ and $k = \frac{R_3}{2R_2}$. The comparator gives a digital (0 or 1) output depending on whether the output of the operational amplifier is positive or negative. This signal controls the driver transistors through the digital delay, which limits the switching frequency of the transistors, so that the overall efficiency remains high. The parameter β should be chosen to give a good transient response to changes in load. The easy extension of this method of implementation to the fourth-order case should be clear. Note that when a load resistance is present we are more concerned with $\dot{V}_0$ than with I_L, so there is no need to measure inductor current.

In practice the operational amplifier will involve a lag term in the measure of $(V_0-\alpha E)$, so that the voltage at its output is more accurately given by $k(\frac{1+\beta s}{1+\delta s})(V_0-\alpha E)$, where $\delta \ll \beta$. Alternatively, it may be necessary to intentionally introduce this lag $\frac{1}{1+\delta s}$ by placing a capacitor C_3 in parallel with R_3, where $C_3R_3 = \delta$, so that the operational amplifier network does not become too receptive to high frequency noise. Although we have not included this lag in our calculations above, it does not affect our conclusions concerning stability (see Ref. 7). The important point, however, is that since the lag term $(\delta s+1)$ will always be present in practice, some form of phase advance $(\beta s+1)$ is necessary for stability, with $\beta > \delta$. Indeed, an analog computer simulation shows that if β is reduced below δ, an oscillation will occur, at a frequency of the order of $\frac{1}{2\pi\sqrt{LC}}$.

ACKNOWLEDGEMENTS

This paper is based on a doctoral dissertation at the Massachusetts Institute of Technology. For help and guidance in this work the author is grateful to Professor Jan C. Willems of the University of Groningen, the Netherlands, and to Professor Roger W. Brockett of Harvard University.

REFERENCES

1. R.W. Brockett and J.L. Willems, "Frequency Domain Stability Criteria: Part I", IEEE Trans. Automatic Control, Vol. AC-10, pp. 255-261, 1965.
2. P.L. Falb and G. Zames, "On Cross-Correlation Bounds and the Positivity of Certain Nonlinear Operators" and "On the Stability of Systems with Monotone and Odd-Monotone Nonlinearities", IEEE Trans. on Automatic Control, Vol. AC-12, pp. 219-223, April 1967.
3. R.P. O'Shea, "A Combined Frequency-Time Domain Stability Criterion for Autonomous Continuous Systems", IEEE Trans. on Automatic Control, Vol. AC-11, pp. 477-483, July 1966; Proc. JACC, pp. 832-840, 1966.
4. J.C. Willems, The Analysis of Feedback Systems, M.I.T. Press, Cambridge, Mass., 1971.
5. J.C. Willems, "Dissipative Dynamical Systems", Archive for Rational Mechanics and Analysis, Volume 45, No. 5, 1972, pp. 321-393.
6. J.L. Willems, Stability Theory of Dynamical Systems, Nelson, London, 1970.
7. J.R. Wood, Power Conversion in Electrical Networks, Ph. D. Thesis, M.I.T., Cambridge, Mass., 1973.

8. G. Zames, "On the Input-Output Stability of Nonlinear Feedback Systems", IEEE Trans. on Automatic Control, Vol. AC-11, pp. 228-238 and 465-476, April and July 1966.

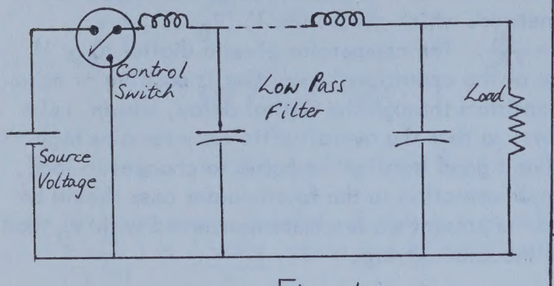

Control Switch
Low Pass Filter
Load
Source Voltage

Fig 1

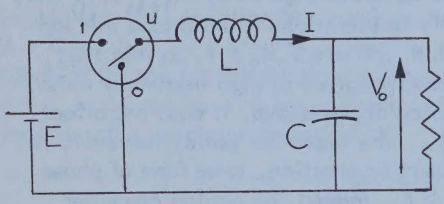

1
u
0
L
I
E
C
V_o
R

Fig 2

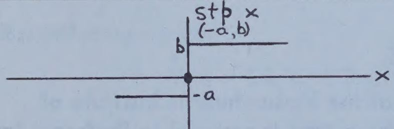


Fig 3

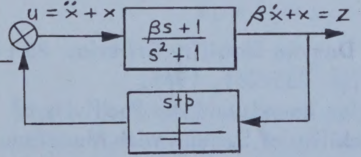

$u = \ddot{x} + x$
$\frac{\beta s+1}{s^2+1}$
$\beta \dot{x} + x = z$
−
s+p

Fig 4

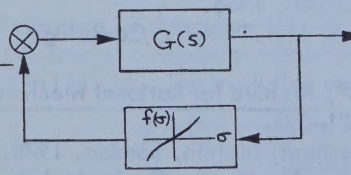


Fig 5

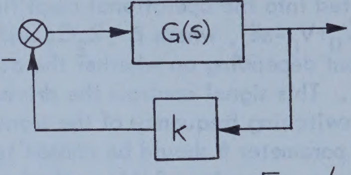


Fig 6

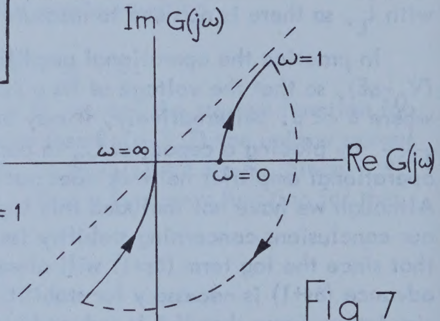


Fig 7

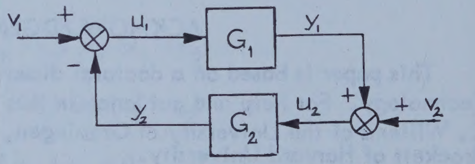

v_1
+
u_1
G_1
y_1
−
+
u_2
+ v_2
y_2
G_2

Fig 8

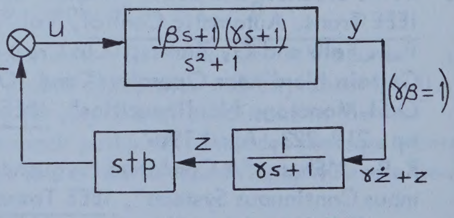

u
$\frac{(\beta s+1)(\gamma s+1)}{s^2+1}$
y
−
$(\gamma\beta = 1)$
z
s+p
$\frac{1}{\gamma s+1}$
$\gamma \dot{z} + z$

Fig 9

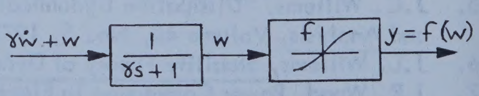

$\gamma \dot{w} + w$
$\frac{1}{\gamma s+1}$
w
f
y = f(w)

Fig 10

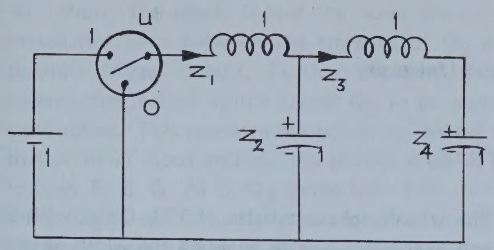

1
u
1
1
z₁
z₃
O
z₂
z₄
+
-
1
1

Fig 11

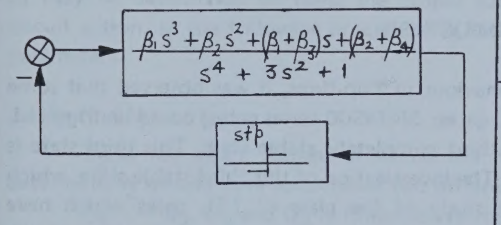

$\frac{\beta_1 s^3 + \beta_2 s^2 + (\beta_1+\beta_3)s + (\beta_2+\beta_4)}{s^4 + 3s^2 + 1}$
-
s+p

Fig 12

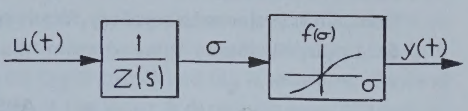

u(t)
1/Z(s)
σ
f(σ)
σ
y(t)

Fig 13

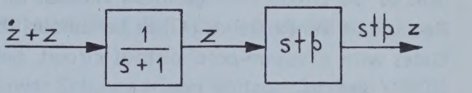

ż+z
1/(s+1)
z
s+p
s+p z

Fig 14

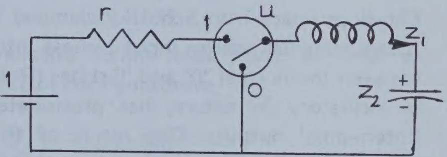


Fig 15

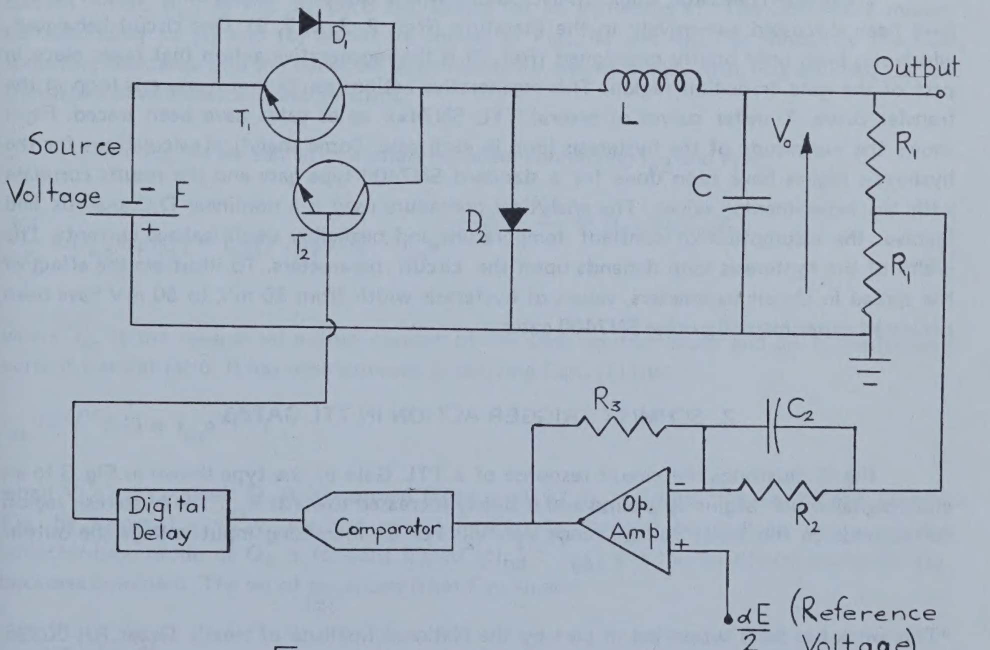

D₁
T₁
T₂
Source Voltage
- E
+
L
Output
V₀
R₁
C
R₁
D₂
R₃
C₂
Op. Amp
-
+
R₂
Comparator
Digital Delay
αE/2 (Reference Voltage)

Fig 16

ON THE HYSTERESIS OF THE TTL CIRCUIT WITH TOTEM-POLE OUTPUT*

R. SRINIVASAN and T. J. CHANEY
Computer Systems Laboratory, Washington University
St. Louis, Missouri

ABSTRACT

The existence of a hysteresis loop in the transfer characteristic of TTL Gates with a 'totem-pole' output is described. A detailed analytical procedure to evaluate the width of the hysteresis loop is explained. Further, a theoretical analysis is followed to prove that the loop gain of the circuit configuration exceeds unity which is the condition for regenerative action. Because of the existence of this behaviour, the input characteristics of some of the TTL NAND Gates with a 'totem-pole' output circuit behave like Current-controlled Negative Resistance (CNR) devices.

I. INTRODUCTION

While investigating the metastable behaviour of Flip-flops, it was observed that some Flip-flops (made from Schottky clamped TTL gates; SN74S00 series gates) could be triggered, (using marginal energy input pulses) into a third completely stable state. This third state is between the normal '0' and 'I' states (Ref. I). The investigation of this third stable state, which is oscillatory in nature, has precipitated a study of the class of TTL gates which have 'totem-pole' outputs. One result of this investigation has been a detailed study of the hysteresis, or Schmitt trigger action, that is inherent in this class of TTL gates.

Transistor-Transistor Logic NAND Gates which utilize a 'totem-pole' output circuit have been discussed extensively in the literature (Ref. 2, 3, 4, 7, 9). One circuit behaviour, which has been only briefly mentioned (Ref. 2) is the regenerative action that takes place in part of the gate transition region. This regenerative action results in a hysteresis loop in the transfer curve. Transfer curves of several TTL SN74xx series gates have been traced. Fig. I shows the magnitude of the hysteresis loop in each case. Some analytical calculations for the hysteresis region have been done for a standard SN7400 type gate and the results correlate with the experimental values. The analytical procedure used is a nonlinear D.C. analysis and involves the assumption of constant temperature and negligibly small leakage currents. The width of the hysteresis loop depends upon the circuit parameters. To illustrate the effect of the spread in circuit parameters, values of hysteresis width from 30 mV to 50 mV have been measured experimentally using SN7400 gates.

2. SCHMITT TRIGGER ACTION IN TTL GATES

Fig. 2 illustrates the circuit response of a TTL Gate of the type shown in Fig. 3 to an input signal which begins at ground and is slowly increased towards V_{CC}. The hysteresis region corresponds to the input voltage range V_1 - V_2. For an increasing input voltage the output

---

*This work has been supported in part by the National Institute of Heath Grant RR-00396 from the Division of Research Resources and in part by the Advanced Research Projects Agency of the Department of Defense under contract SD-302.

follows the path A, B, C, D. At D, transistors Q_2 and Q_3 are in linear active region and Q_4 is in saturation. The diode D and the now forward-biased collector-base diode of Q_4 form a low impedance path between the emitters of Q_2 and Q_3. The circuit may now be viewed as a Schmitt trigger circuit. Further increase of the input voltage above point D results in regenerative action which causes Q_3 to go more and more towards cutoff and Q_2 more into conduction. This regenerative action continues until Q_3 is cutoff and Q_2 is saturated. Beyond this point in input voltage, the output is along E, F. If the input is decreased now, the output follows F, E, G. At G, Q_3 comes back into conduction and Q_2 enters the linear active region. Since Q_4 is still saturated the regenerative action starts again and the output jumps to H. Further decrease of input voltage makes the output follow the path C, B, A.

3. ANALYTICAL PROCEDURE TO DETERMINE THE WIDTH OF HYSTERESIS

A theoretical evaluation of the upper and lower Schmitt trigger voltages V_1 and V_2 (Fig. 2) may be determined by using the proper Ebers-Moll models (Ref. 8) for each region of circuit action. In the following procedure V_1 and V_2 are evaluated for a particular set of circuit parameters.

3.1 To find V_2:

V_2 is the input voltage at which the regenerative action takes place. In order to determine V_2 we will have to consider two different circuit configurations:

(a) Q_2, Q_3 and Q_4 in linear active region
(b) Q_2 and Q_3 in linear active region and Q_4 in saturation.

The aim is to find that value of input voltage corresponding to which the emitter current of Q_3 (I_{e3} in Fig. 3) starts decreasing for an increasing V_i. The Ebers-Moll models corresponding to (a) and (b) above are shown in Figs. 4a and 4b. For simplicity the base spreading resistance and the collector series resistance are neglected. Also, it is assumed that all transistors have identical characteristics.

From Fig. 4a we can write a single equation connecting I_{e3} and V_i as:

$$\frac{KT}{qR_3}\ln[I_{e3}/\alpha_F I_{es}] + I_{e3}(1-\alpha_F)/\alpha_F = I_{es}e^{\frac{q}{KT}\left[V_i - \frac{KT}{q}\ln(I_{e3}/\alpha_F I_{es})\right]} \tag{1}$$

where I_{es} is the reverse saturation current of the base emitter diode and α_F is the forward current transfer ratio. It has been assumed in deriving Eqn. (1) that

$$I_{es}\,(e^{V/V_T}-1) \cong I_{es}e^{V/V_T}\ .$$

when $V >> 5V_T$ (Ref. 5). It is assumed that $\alpha_F = 0.97$, $I_{es} = 2 \times 10^{-15}$A and $KT/q = 26$ mV. Fig. 5a shows I_{e3} vs. v_i for the model in Fig. 4a. Beyond point P on this curve the collector-base diode of Q_4 is forward biased ($\beta I_{b4} > I_{c4}$) and the saturation model for Q_4 becomes dominant. The set of equations from Fig. 4b are:

$$I_{es}e^{\frac{q}{KT}(V_i - V_e)} = V_e/R_3 + I_{es}\,(1-\alpha_F)\,e^{qV_e/KT} + I_{cs}\,(1-\alpha_R)\,e^{q(V_e - V_o)/KT} \tag{2}$$

$$V_{cc} - \alpha_F\ I_{e2}R_2 - V_o = \frac{2KT}{q}\ \ln\frac{I_{e3}}{I_{es}} + I_{e3}(1 - \alpha_F)R_2 \qquad (3)$$

where I_{cs} is the reverse saturation current of collector-base diode and α_R is the reverse current amplification factor. I_{cs}, calculated from I_{es} with $\alpha_R = 0.1$, is 19.4×10^{-15}A. From Eqn. (2),

$$V_o = \frac{1}{40}\ \ln\left[\frac{I_{cs}(1 - \alpha_R)\ e^{qV_e/KT}}{I_{es}e^{q(V_i - V_e)/KT} - V_e/R_3 - I_{es}(1 - \alpha_F)\ e^{qV_e/KT}}\right] \qquad (4)$$

Egns. (2), (3) and (4) are solved using iterative methods in order to accurately evaluate the inflection point on the I_{e3} vs. V_i curve. Fig. 5b shows only the region of interest where I_{e3} starts decreasing. The inflection point in Fig. 5b corresponds to the downswitching voltage v_2 (Fig. 2). For the assumed set of circuit parameters $V_2 = 1.434$V and the corresponding output voltage $V_o = 0.236$V.

3.2 To find V_1:

At the end of the transition region DE (Fig. 2), Q_3 is cutoff, Q_2 and Q_4 are saturated. The output voltage V_0 stays constant along EF for all $V_i > V_2$. In order to find V_1 the input voltage has to be decreased along F, E, G to determine the point where Q_3 enters the linear active region once again i.e. when V_{Be3} reaches the cut in voltage of approximately 0.5V (Ref. 6). Before reaching the upper switching point G (Fig. 2) the circuit configurations to be considered are:

(a) Q_2 and Q_4 in saturation and Q_3 cutoff and

(b) Q_2 in linear active region, Q_3 cutoff and Q_4 saturated.

The Ebers-Moll model for (a) above is shown in Fig. 6 and from this we get:

$$I_{es}(1 - \alpha_F\alpha_R)\ e^{\frac{qV_i}{KT}} = e^{\frac{qV_e}{KT}}\left[V_e/R_3 + I_{es}\ (1 - \alpha_F\alpha_R)\ e^{\frac{qV_e}{KT}} - \alpha_R(V_{cc}-V_{c2})/R_2\right] \qquad (5)$$

V_e is evaluated for different values of $V_i \leqslant V_2$ and for each V_e it is possible to find the corresponding value of V_0 from Eqn. (4) and hence $v_{Be3} \cong (V_{c2} - V_0)/2$.

It has been found that below a particular value of V_i the linear active model of Q_2 becomes dominant. The circuit configuration corresponding to (b) is represented by the model similar to that in Fig. 6 excepting that I_{R2} and $\alpha_R I_{R2}$ are omitted as $I_{R2} << \alpha_F I_{F2}$ and $\alpha_R I_{R2} << I_{F2}$. The resulting equation obtained is:

$$I_{es}e^{qV_i/KT} = e^{qV_e/KT}\left[V_2/R_3 + I_{es}\ (1 - \alpha_F\alpha_R)\ e^{qV_e/KT}\right] \qquad (6)$$

From the known values of V_i, V_e could be found and hence the corresponding values of V_0 from Eqn. (4). V_{c2} is computed from the following equations.

$$V_{Be3} \cong (V_{c2} - V_o)/2 \qquad (7)$$

$$I_{e2} \cong I_{es}\, e^{q(V_i - V_e)/KT} \tag{8}$$

$$I_{R1} \cong I_{cs}\, e^{q(V_e - V_o)/KT} \tag{9}$$

$$I_{R1} \cong \alpha_F I_{FI} \cong \alpha_F\, I_{es}\, e^{qV_e/KT} \tag{10}$$

$$V_{c2} = V_{cc} - \alpha_F\, I_{e2}\, R_2 \tag{11}$$

That value of V_i corresponding to $V_{Be3} = 0.5V$ is the upswitching voltage V_1 in Fig. 2. For the assumed values of circuit parameters,

$V_1 = 1.388V$

$V_2 = 1.434V$

and hence the width of hysteresis = 46 mV.

4. CALCULATION OF LOOP GAIN

The discussed regenerative action will take place only if the internal loop gain is greater than unity. In order to calculate the loop gain, Q_4 (in saturation) is characterized by a two-port network and Q_2 and Q_3 by their small signal modes as shown in Fig. 7 (Ref. 10). The input transistor Q_1 behaves like two forward biased diodes (CB and BE) in this region of regenerative action. Z is the impedance looking from the base of Q_2 towards the collector of Q_1. Neglecting the resistance of forward biased diodes of Q_1, $Z \cong R_1 \parallel R_s$ where R_s is the source impedance at the input.

$$\begin{aligned} I_1 &= y_{11}V_1 + y_{12}V_2 \\ I_2 &= y_{21}V_1 + y_{22}V_2 \end{aligned} \tag{12}$$

From the Ebers-Moll model of Q_4 and Fig. 7, it is found that :

$$\begin{aligned} &y_{11} = I_R(q/KT),\ y_{12} = I_c(q/KT),\ y_{21} = -I_R\,(q/KT)\,(1 - \alpha_R), \\ &y_{22} = I_b(q/KT),\ V_1 = V_{ce} \text{ and } V_2 = V_e \end{aligned} \tag{13}$$

$$\begin{aligned} &[y_{11} + (1 + g_{m2}r_{\pi 2})/r_{\pi 2}]\ V_1 + y_{12}V_2 = [(1 + g_{m2}r_{\pi 2})/r_{\pi 2}]V \\ &y_{21}V_1 + [y_{22} + (1 + g_{ml}r_{\pi 1})/(Z + r_{\pi 1}) + 1/R_3]V_2 = 0 \end{aligned} \tag{14}$$

$$\text{Loop Gain} = V'/V = \frac{K_4K_5}{K_3K_5 + K_4K_6 - K_1K_7} \tag{15}$$

where $K_1 = y_{11} + (1 + g_{m2}r_{\pi 2})/r_{\pi 2}$

$K_2 = y_{12}\,(Z + r_{\pi 1})/g_{m1}r_{\pi 1}r_{\pi 2}$

$K_3 = K_2[1 + (r_{\pi 2}/R_2)]$

$K_4 = (1 + g_{m2}r_{\pi 2})/r_{\pi 2}$

$K_5 = y_{21}$

$K_6 = [y_{22} + (1 + g_{m1}r_{\pi 1})/(Z + r_{\pi 1}) + 1/R_3]\ [Z + r_{\pi 1})/\ g_{m1}r_{\pi 1}r_{\pi 2}]$

$K_7 = K_6[1 + (r_{\pi 2}/R_2)]$

A computer program has been written to evaluate the two-port parameters for each D.C. condition in the regenerative region for a decreasing I_{e3} and the loop gain has been calculated in each case. A plot of Loop Gain vs the Emitter current of Q_3 ($=I_{e3}$) is shown in Fig. 8. For this typical curve Z has been assumed to be 50 Ohms since R_s in the experimental set up is 50 ohms and $Z \cong R_1 || R_s$. $R_1 = 4.7K$ and hence $Z \cong 50$ ohms. The loop gain is greater than unity for I_{e3} less than 21 mA and greater than a few micro amperes. This region of decreasing I_{e3} corresponds to the downswitching schmitt trigger voltage. By a similar analysis the loop gain corresponding to the upswitching region could be computed.

5. INPUT CHARACTERISTIC

The input characteristic of a typical SN7400, Fig. 9, shows a negative resistance region which is not discussed by the manufacturers of IC TTL. SN7400 type gates can be viewed as Current-controlled Negative Resistance Devices (CNR) over a range of input voltage. One possible application of this characteristic is that the SN7400's could be used to produce relaxation oscillations as shown in Fig. 10.

The regenerative action in the Schottky TTL SN74S00 is responsible for producing a transfer characteristic of two gates in cascade as shown in Fig. 11a. This characteristic is similar to an ideal Schmitt trigger. The response of this cascade circuit to a 100 Hz. sine wave is shown in Fig. 11b. The rise and fall time of this square wave are 60 n.sec. and 80 n.sec. respectively.

6. CONCLUSION

The existence of Schmitt trigger action in TTL gates with a totem-pole output has been described. Values of the hysteresis width obtained from theoretical and experimental methods agree very well. The upswitching and downswitching Schmitt trigger voltages depend upon the circuit parameters and hence the hysteresis width varies from one gate to the other in the same family and from one family to the other. The input characteristic of SN7400 type gates show that they behave like Current-controlled Negative Resistance devices.

ACKNOWLEDGMENTS

We wish to thank Dr. Charles E. Molnar and Dr. Robert O. Gregory for many useful discussions and encouragement. Our sincere thanks are also due to Dr. Levente Mate for writing the computer algorithm and valuable discussions.

REFERENCES

1. C.L. Seitz: Personal Communication, Computer Science Department, The University of Utah, Salt Lake City, Utah, June, 1972.

2. D.K. Lynn, C.S. Meyer and D.J. Hamilton: ‘Analysis and Design of Integrated Circuits’, McGraw-Hill Book Company, New York, 1967, pp. 312-313.

3. P.J. Tizzard and M.J. Turner: ‘Analysis of TTL Operation and Design’, Microelectronics, Jan. 1968, Vol. 1, pp. 28-33.

4. N.B. Rabbat, et. al.: ‘Transient Analysis of Multiple-Input Integrated Digital Structures’, IEEE Transactions on Electron Series, Vol. ED 20, No. 5, pp. 496-508, May, 1973.

5. L. Strauss: ‘Wave Generation and Shaping’, McGraw-Hill Book Company, New York, 1960, p. 51.

6. J.M. Millman and Taub: ‘Pulse, Digital, and Switching Wave-forms’, McGraw-Hill Book Company, New York, 1965, p. 217.

7. Yasuo Tarui, et. al.: ‘Transistor Schottky-Barrier Diode Integrated Logic Circuit’, IEEE Solid State Circuits, Vol. SC-4, No. 1, pp. 3-12, February 1969.

8. Ebers, J.J., and J.L. Moll: ‘Large-signal Behaviour of Junction Transistors’, Proc. IRE, Vol. 42, pp. 1761-1772, December 1954.

9. John A. DeFalco: ‘Comparison and Uses of TTL Circuits’, Computer Design, Vol. 11, No. 2, pp. 63-68, February, 1972.

10. J.N. Harris, P.E. Gray and C.L. Searle: ‘Digital Transistor Circuits’, John Wiley & Sons, New York, 1966.

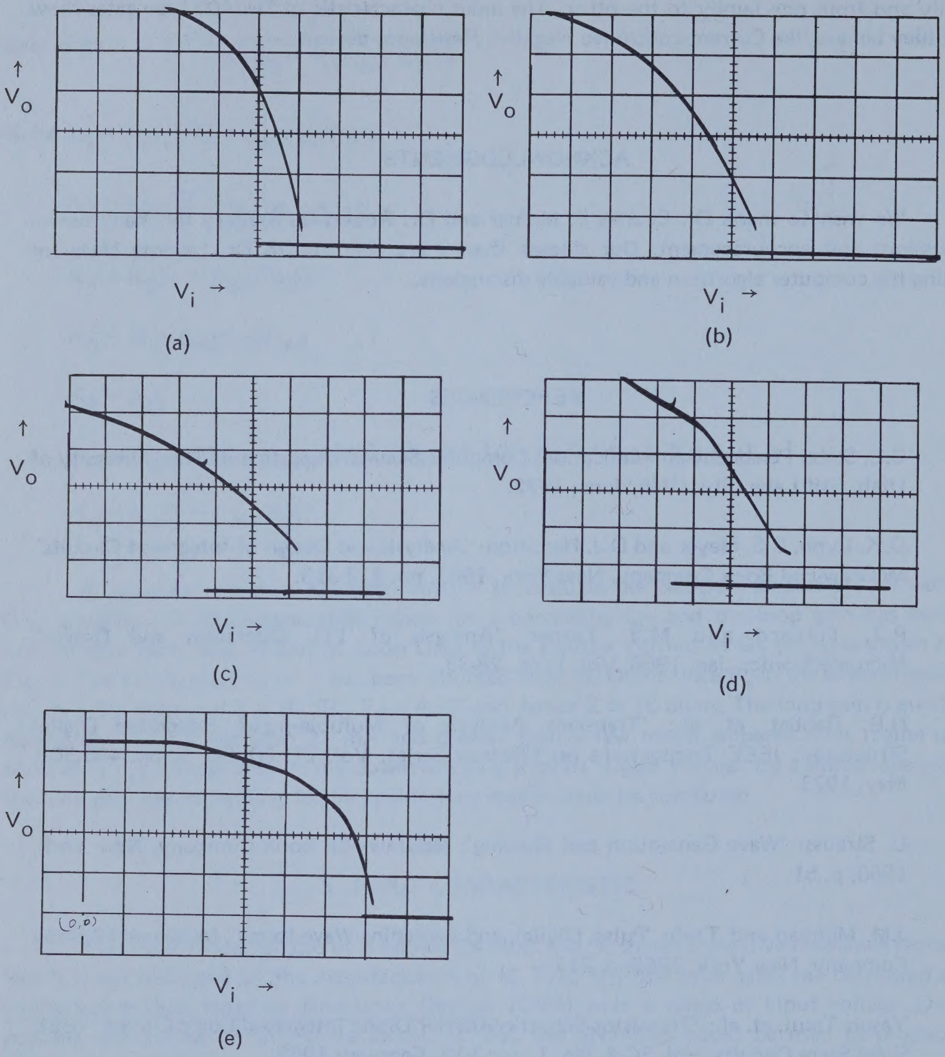


Fig. 1.

Hysteresis Region of the transfer characteristics of Low Power, Standard, High Power and Schottky-clamped TTL type Gates.

a. SN74L00
b. SN7400
c. SN74H00
d. SN74S00

Scale: V_i @ 50mV/div. V_o @ 0.5V/div.

e. Complete transfer characteristic of SN7400 (origin noted on photo)

Scale: V_i @ 0.2V/div. V_o @ 1.0V/div.

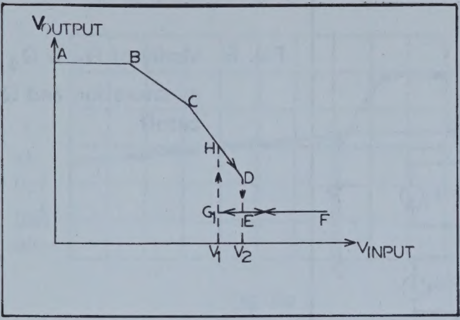


Fig. 2
Transfer Characteristic
of SN7400

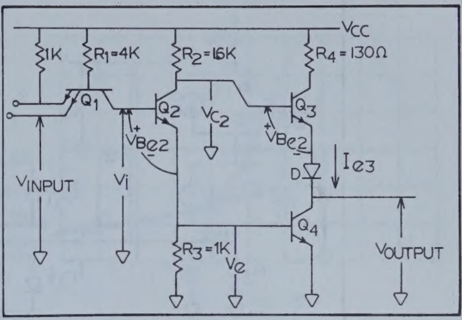

VCC
1K
R1=4K
R2=16K
R4=130Ω
Q1
Q2
VC2
Q3
VBe2
VBe2
Ie3
D
VINPUT
Vi
Q4
R3=1K
Ve
VOUTPUT

Fig. 3
Schematic of ¼SN7400
(with one input active)

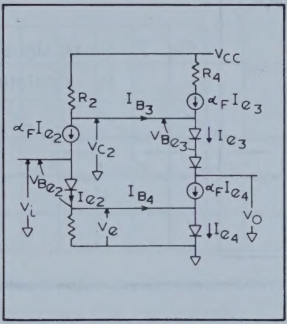

VCC
R4
R2
IB3
αFIe3
αFIe2
VC2
VBe3
Ie3
VBe2
Ie2
IB4
αFIe4
Vi
Ve
VO
Ie4

Fig. 4a
Model of Q_2, Q_3 & Q_4
in Linear Active Region

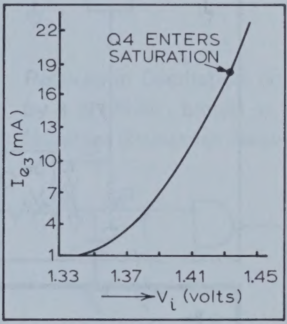


Fig. 5a
I_{e3} vs. V_i from the
model in Fig. 4a.

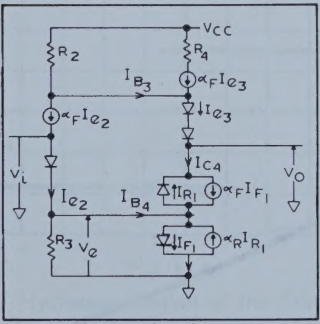

VCC
R2
R4
IB3
αFIe3
αFIe2
Ie3
IC4
Vi
VO
IR1
αFIF1
Ie2
IB4
R3
Ve
IF1
αRIR1

Fig. 4b
Model of Q_2 & Q_3
in Linear Active Region &
Q_4 in Saturation

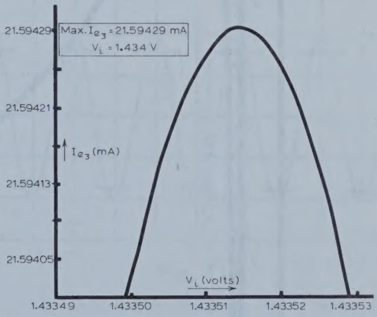


Fig. 5b
I_{e3} vs. V_i from the model in
Fig. 4b.

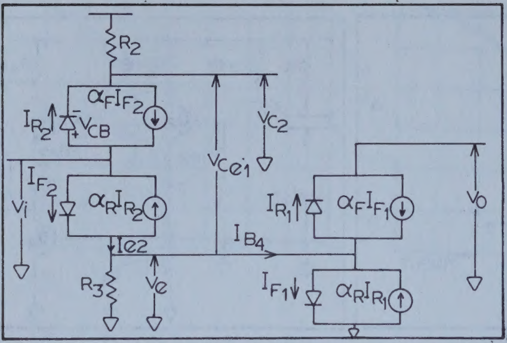

R2
αF IF2
IR2
VCB
VC2
VCe1
Vi
IF2
αR IR2
IR1
αF IF1
Vo
IB4
Ie2
R3
Ve
IF1
αR IR1

Fig. 6. Model of Q_2 & Q_4 in saturation and Q_3 cutoff.

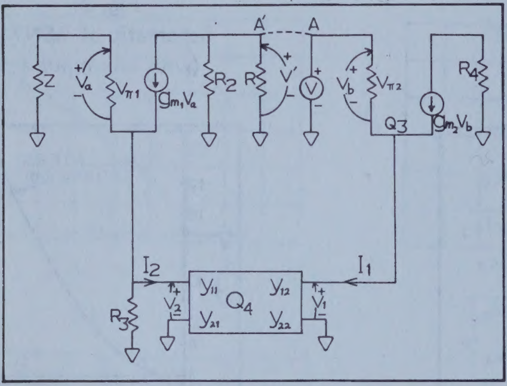

A'
A
Z
Va
Vπ1
gm1Va
R2
R
V'
V
Vb
Vπ2
R4
Q3
gm2Vb
I2
I1
y11
y12
Q4
y21
y22
V2
V1
R3

Fig. 7. Small signal model to calculate loop gain.

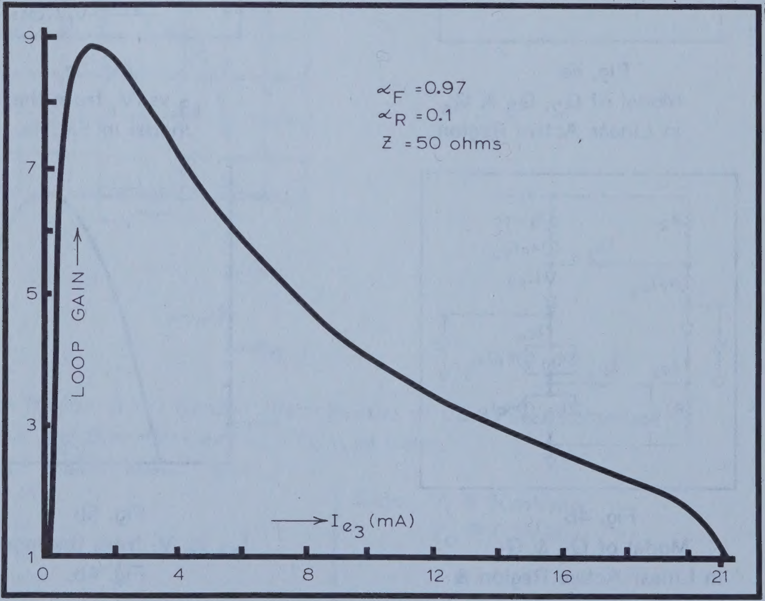


Fig. 8
Loop Gain vs. I_{e3}

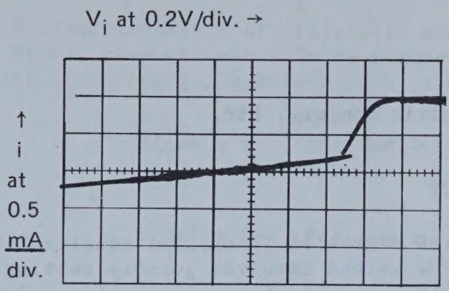


Fig. 9a
Input Characteristic
of SN7400

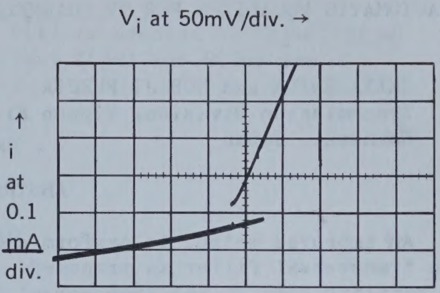
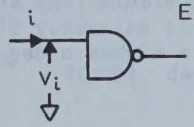
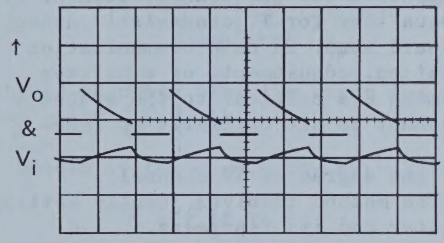


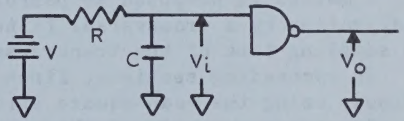
Fig. 9b
Expanded Hysteresis Region





time at 0.1 m. sec/div. →
Top trace: V_o at 1V/div
Bottom trace: V_i at 0.1V/div.

Fig. 10. Relaxation Oscillation produced by a SN7400 - biased in the Negative Resistance Region:



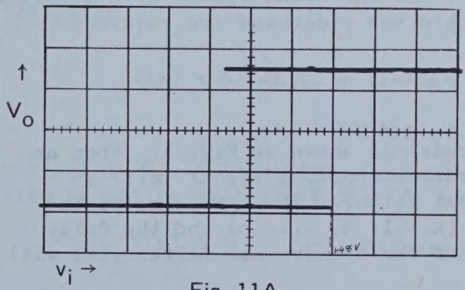


Fig. 11A
Hysteresis Region of the Transfer Characteristic of. two SN74S00's connected in cascade.
Scale: V_o at 1V/div
V_i at 50 mV/div

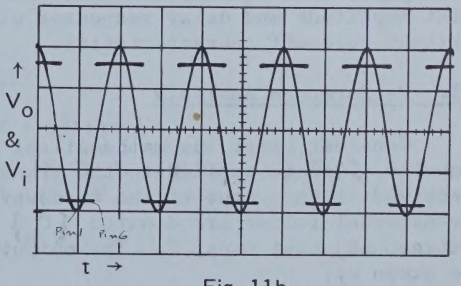


Fig. 11b
Square-wave output, for a sinusoidal input, from the cascade.
Scale: V_o & V_i at lV/div
time, t, at 5 m.sec/div

AN AUTOMATIC EQUALIZER FOR TV CHANNELS

SEIYA SHIDA and NOBUJI FURUYA
Transmission Division, Nippon Electric Company, Ltd.
Kawasaki, Japan

ABSTRACT

An improved method to perform a fine degree of TV channel equalization by a transversal filter is proposed. The method involves jointly setting the sampling time of the transversal filter and its tap gains, using a Zero Forcing method. The effectiveness of this method is shown theoretically. It is shown, by both computer simulations and experiments, that the equalized frequency response is extremely flat compared with that by the usual equalizer, especially in higher order group delay distortions.

I. Introduction

A fine degree of equalization is required for the transmission of TV signals. E. Arnon showed an adaptive equalizer for TV channels(1) using Lucky's Zero Forcing method(2). It is said that, in data communication systems with transversal filter equalization, adjustments of equalizer taps, carrier phase and the receiver timing are critical to the system's performance and that these parameters cannot be set optimally by independent adjustments(3).

A method is proposed to perform a fine degree of TV channel equalization by a transversal filter. The method involves jointly setting the sampling time of the transversal filter and its tap gains.

In succeeding sections, first the error function to be minimized is defined, using the mean-square error equalization criterion and a few considerations are made of the error function. Applying the considerations, a simple Zero Forcing equalizer was constructed which includes sampling time control function. Next, results obtained with the experiments are compared with those obtained by simulations. The comparison is made mainly in the frequency domain because, for the transmission of a TV signal, flat amplitude and delay responses within the videoband are required without roll-off characteristics.

II. Performance Analysis

Consider first the mathematical model as shown in Fig. 1. When an impulse $\delta(t)$ is applied to the transmitter input, a signal $x(t)$ is received at the input of the transversal filter. The tap gains of the transversal filter are denoted $\{C_k\}$ $(k = 1, 2, \ldots, n)$ and the delay between adjacent taps, T. The output of the transversal filter $y(t)$ will be given by,

$$y(t) = \sum_{k=1}^{n} c_k x(t-kT) \qquad (1)$$

The output of the transversal filter is delayed τ second through the ideal delay line, which has the physical meaning of the receiver sampling time control, and is compared to the reference signal $r(t)$.

We define error function $\mathcal{E}$ to be minimized, using the mean-square error equalization criterion, i.e.,

$$\mathcal{E} = \int_{-\infty}^{\infty} \left\{ y(t-\tau) - r(t) \right\}^2 dt \qquad (2)$$

The Fourier transform of x(t), y(t) and r(t) is denoted as $X(j\omega)$, $Y(j\omega)$ and $R(j\omega)$, respectively. It is assumed that $X(j\omega)$ and $R(j\omega)$ are bandlimitted, where bandwidth is W, i.e.,

$$X(j\omega) = 0 \quad \text{for } |\omega| > 2\pi W \tag{3}$$

$$R(j\omega) = 0 \quad \text{for } |\omega| > 2\pi W \tag{4}$$

Using Perseval's theorem, Eq. (2) is rewritten as,

$$\begin{aligned} \mathcal{E} &= \int_{-2\pi W}^{2\pi W} \left| Y(j\omega)e^{-j\omega\tau} - R(j\omega)\right|^2 d\omega \\ &= \int_{-2\pi W}^{2\pi W} \left| \sum_{k=1}^{n} C_k X(j\omega)e^{-j\omega(kT+\tau)} - R(j\omega)\right|^2 d\omega \end{aligned} \tag{5}$$

It should be noted that the error function in (5) is not only the function of tap gains $\{C_k\}$ (k = 1, 2, ..., n), but also the function of sampling time τ.

Denote it for simplicity by,

$$\mathcal{E} \equiv \mathcal{E}(\mathbb{C},\tau) \tag{6}$$

, where $\mathbb{C}$ means the tap gains vector.

The error function to be minimized is defined by (6). A theoretical study is carried out to determine whether error function $\mathcal{E}$ will converge to the absolute minimum when parameters $\mathbb{C}$ and τ are set by the gradient method. For this purpose, consider function $P(\lambda)$, which is defined by the following,

$$\begin{aligned} P(\lambda) &= \lambda\mathcal{E}(\mathbb{C},\tau) + (1-\lambda)\mathcal{E}(\mathbb{C}',\tau') \\ &\quad - \mathcal{E}(\lambda\mathbb{C} + (1-\lambda)\mathbb{C}', \lambda\tau + (1-\lambda)\tau') \end{aligned} \tag{7}$$

, where point $(\mathbb{C}', \tau')$ differs from point $(\mathbb{C}, \tau)$. $\mathcal{E}$ is a convex function of the parameters $(\mathbb{C},\tau)$ if, and only if,

$$P(\lambda) \geqq 0 \quad \text{for arbitrary } \lambda \ (0 \leqq \lambda \leqq 1) \tag{8}$$

It is shown that $\mathcal{E}$ is not a convex function of the parameters $(\mathbb{C},\tau)$ by showing one counter-example (see the Appendix). Therefore, the following proposition about convergence can be obtained.

Proposition 1

When equalizer tap gains $\{C_k\}$ (k = 1, 2, ..., n) and sampling time τ are jointly set by the gradient method, equalization error $\mathcal{E}$ will converge to one of the local minimums.

Consider next the relation between error function and output signal of the transversal filter. It is assumed that the reference characteristic satisfies the following equation, i.e.,

$$R(j\omega) = 1 \qquad \text{for } |\omega| \leqq 2\pi W \tag{9}$$

We denote $Y(j\omega)$ in terms of the amplitude, $A(\omega)$, and phase, $\beta(\omega)$, i.e.,

$$Y(j\omega) = A(\omega)e^{j\beta(\omega)} \tag{10}$$

, where the flat delay component is removed from $\beta(\omega)$.

Output $y(t)$ and error ε of Eq. (5) are given in the following, respectively,

$$y(t) = \frac{1}{2\pi}\int_{-2\pi W}^{2\pi W} A(\omega)e^{j(\beta(\omega) + \omega t)}\, d\omega \tag{11}$$

$$\varepsilon(\mathbb{C},\tau) = \int_{-2\pi W}^{2\pi W} \left| A(\omega)e^{j(\beta(\omega) - \omega\tau)} - 1\right|^2 d\omega \tag{12}$$

Assuming that $A(\omega)$ as an even function and $\beta(\omega)$ as an odd function, we can get the following equations, i.e.,

$$\frac{dy(t)}{dt} = -\frac{1}{\pi}\int_0^{2\pi W} \omega A(\omega)\sin(\beta(\omega) + \omega t)\, d\omega \tag{13}$$

$$\frac{d^2y(t)}{dt^2} = -\frac{1}{\pi}\int_0^{2\pi W} \omega^2 A(\omega)\cos(\beta(\omega) + \omega t)\, d\omega \tag{14}$$

$$\frac{\partial\varepsilon(\mathbb{C},\tau)}{\partial\tau} = -4\int_0^{2\pi W} \omega A(\omega)\sin(\beta(\omega) - \omega\tau)\, d\omega \tag{15}$$

$$\frac{\partial^2\varepsilon(\mathbb{C},\tau)}{\partial\tau^2} = 4\int_0^{2\pi W} \omega^2 A(\omega)\cos(\beta(\omega) - \omega\tau)\, d\omega \tag{16}$$

Denoting output signal peak time, t_0, then,

$$\left.\frac{dy(t)}{dt}\right|_{t=t_0} = 0 \tag{17}$$

$$\left.\frac{d^2y(t)}{dt^2}\right|_{t=t_0} < 0 \tag{18}$$

If we take $\tau = -t_0$, comparing Eq. (15) to Eqs. (13) and (17), Eq. (16) to Eqs. (14) and (18), we can get the following equations, i.e.,

$$\left.\frac{\partial\varepsilon(\mathbb{C},\tau)}{\partial\tau}\right|_{\tau=-t_0} = 0 \tag{19}$$

$$\left.\frac{\partial^2 \mathcal{E}(\mathbb{C},\tau)}{\partial \tau^2}\right|_{\tau=-t_0} > 0 \tag{20}$$

Therefore, the result can be summarized by the following proposition.

Proposition 2

Assuming Eq.(9) holds, for arbitrary tap gains vector $\mathbb{C}$, the peak time of the output of the transversal filter agrees with one of the minimums of the error function $\mathcal{E}(\mathbb{C},\tau)$ concerning sampling time τ.

Applying the above result, a method of equalization is proposed. The method involves the sampling time being reset at the peak of the output signal after each iteration of tap gains by the gradient method.

The relation was also studied between the convergence and the modified amount of the tap gains and the sampling time of each iteration. Because of space limitation, only the results are presented which are shown as the following proposition.

Proposition 3

The sufficient condition that error function $\mathcal{E}(\mathbb{C},\tau)$ converges to one of the local minimums by the proposed method is,

$$\| y(t) - y'(t) \| \geqq \left[4\pi \left.\frac{dy'(t)}{dt}\right|_{t=t_0+\Delta\tau}\right] \cdot \Delta\tau \tag{21}$$

, where $y'(t)$ is the output of the transversal filter, where tap gains vector is $\mathbb{C}'$ and,

$$\| y(t) - y'(t) \| = \int_{-2\pi W}^{2\pi W} \left| Y(j\omega) - Y'(j\omega) \right|^2 d\omega \tag{22}$$

, and $\Delta\tau$ is the modified amount of the sampling time.

III. Computer simulations and experiments

Applying the above considerations, a simple Zero Forcing equalizer was constructed which includes the sampling time control function.

In hardware, a Zero Forcing equalizer is more easily realizable than a Mean Square equalizer because the former needs no correlators. The equalizer operates in the baseband. Sampling time control is accomplished by tracing the peak position of the received test pulse which is transmitted through channels.

A block diagram of the equalizer is shown in Fig. 2. The test pulse, generated in the transmit unit and with the timing information added, is transmitted through the channel. The transversal filter in the receiver has 21 taps. Each tap gain is discretely changed up to ±127 steps. The delay between adjacent taps, as determined by the Nyquist rate, is 0.5 µs. The controller of the equalizer includes the function for tracing the peak position of the received test pulse. This function is accomplished by the pre-amplifier (PRE AMP) and the phase shifter (PS) shown in Fig. 2.

To examine the capability of the equalizer, several distortion experiments were performed and compared with simulations. Two of these are shown in Fig. 3 and Fig. 4. A 6 dB, 0.56 µs slope distortion experiment is shown in Fig. 3. Frequency characteristics of attenuation and delay are, respectively, shown in Fig. 3(a) and Fig. 3(b).

In this case, the distortions were reduced to within 0.5 dB and 0.15 µs, respectively. Computer simulation results are also duplicated. Test pulse waveforms are also shown in Fig. 3(c).

By computer simulation, the method of jointly setting sampling time and tap gains was compared to the earlier method, which only set the tap gains. The results are shown in Figs. 3(d) to (f).

Using a 2.7 dB, 1.34 µs bump distortion experiment, the results by both computer simulations and experiments are also shown in Fig. 4, in the same manner as in Fig. 3.

It is also shown that group delay ripples, after equalization by the proposed method, are several times less than those obtained by earlier method.

IV. Conclusion

Sampling time deviation is sensitive to the characteristics to be equalized. Experiments were made on a Zero Forcing TV channel equalizer which has the function of sampling time control. Experimental results show good performance.

Computer simulation also shows that the group delay ripples after equalization by the proposed method are several times less than those obtained by earlier methods.

V. Acknowledgement

The authors wish to express their thanks to Mr. Tohru Tsuchiya and Mr. Miyoji Ooyama for their kind suggestions and discussions.

VI. Appendix

Substituting Eq.(5) into Eq.(7), P(λ) is rewritten as follows,

$$P(\lambda) = \lambda\int_{-2\pi W}^{2\pi W}\left|Y(j\omega)e^{-j\omega\tau} - R(j\omega)\right|^2 d\omega$$

$$+ (1-\lambda)\int_{-2\pi W}^{2\pi W}\left|Y'(j\omega)e^{-j\omega\tau'} - R(j\omega)\right|^2 d\omega$$

$$- \int_{-2\pi W}^{2\pi W}\left|\left\{\lambda Y(j\omega)+(1-\lambda)Y'(j\omega)\right\} e^{-j\omega\{\lambda\tau+(1-\lambda)\tau'\}} - R(j\omega)\right|^2 d\omega \qquad (23)$$

, where Y'(jω) is the output of the transversal filter whose tap gains vector is $\mathbb{C}'$.

Taking the value in Eq. (22) as follows,

$$\begin{cases} Y(j\omega) = Y'(j\omega) = R(j\omega) = 1 \quad \text{for} |\omega| \leqq 2\pi W \\ \tau = 1/(4W) \quad , \quad \tau' = 5/(4W) \end{cases} \qquad (24)$$

Eq. (22) is rewritten as follows,

$$P(\lambda) = \int_{-2\pi W}^{2\pi W} \left\{ \lambda \left| e^{-j\frac{\omega}{4W}} -1 \right|^2 + (1-\lambda) \left| e^{-j\frac{5\omega}{4W}} - 1 \right|^2 - \left| e^{-j\frac{(5-4\lambda)\omega}{4W}} -1 \right|^2 \right\} d\omega$$

$$= 16W\left(- \frac{1+4\lambda}{5} + \frac{1}{5-4\lambda}\sin \frac{5-4\lambda}{2}\pi\right) \qquad (25)$$

So, taking $\lambda = 1/2$, then,

$$P(\tfrac{1}{2}) = - \frac{16\times 14}{15} W < 0 \qquad (26)$$

This example shows that ε is not a convex function of the parameters (c, τ).

VII. Reference

(1) E. Arnon, "An Adaptive Equalizer for Television Channels," IEEE Trans. on COMTEC Vol. COM-17, No.6, December 1969.

(2) R.W. Lucky, "Automatic Equalization for digital communication," BSTJ Tech. J., vol. 44, pp. 547-588, April 1965.

(3) R.W. Chang, "Joint Equalization, Carrier Acquisition and Timing Recovery for Data Communication," ICC. 70-CP-331-COM

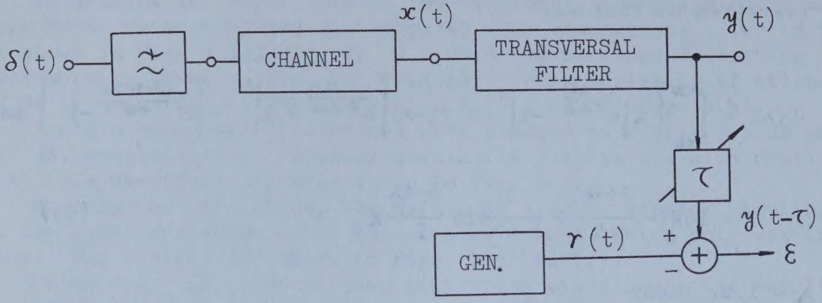
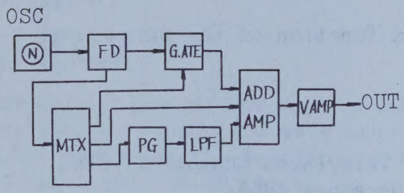

$\delta(t)$
CHANNEL
$x(t)$
TRANSVERSAL FILTER
$y(t)$
τ
$y(t-\tau)$
GEN.
$r(t)$
+
−
ε

Fig. 1. A model of automatic equalization


OSC
N
FD
GATE
MTX
PG
LPF
ADD AMP
VAMP
OUT

FD ; Freq. Divider

PG ; Pulse Generator

VAMP ; Video Amp.

MTX ; Matrix Circuit

REG ; Regenerator

PS ; Phase Shifter

MULT ; Freq. Multiplexer

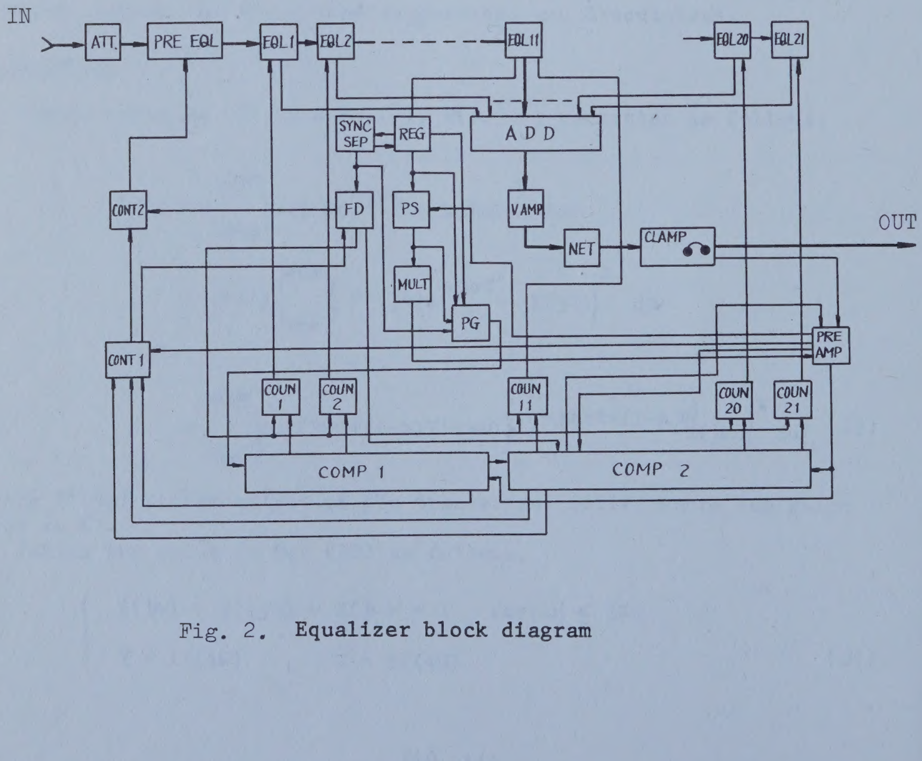
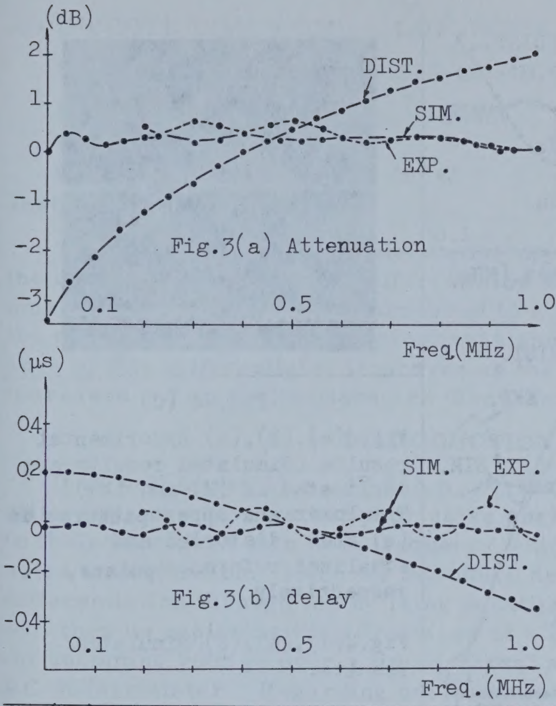

IN
ATT.
PRE EQL
EQL 1
EQL 2
EQL 11
EQL 20
EQL 21
SYNC SEP
REG
ADD
FD
PS
VAMP
NET
CLAMP
OUT
MULT
PG
CONT 2
CONT 1
PRE AMP
COUN 1
COUN 2
COUN 11
COUN 20
COUN 21
COMP 1
COMP 2

Fig. 2. Equalizer block diagram



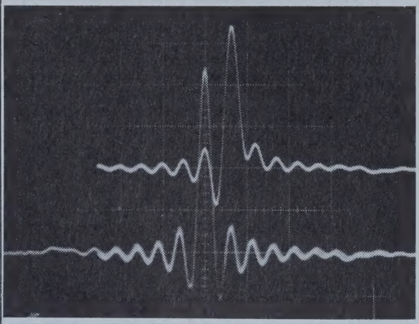


Fig. 3(C)

Fig.3(a),(b),(c) Experimental results (Simulated results are duplicated in (a) and (b)). The upper and lower **pictures in (c)** show distorted and equalized reference **pulses, respectively.**

Fig.3(d),(e),(f) Simulated results.

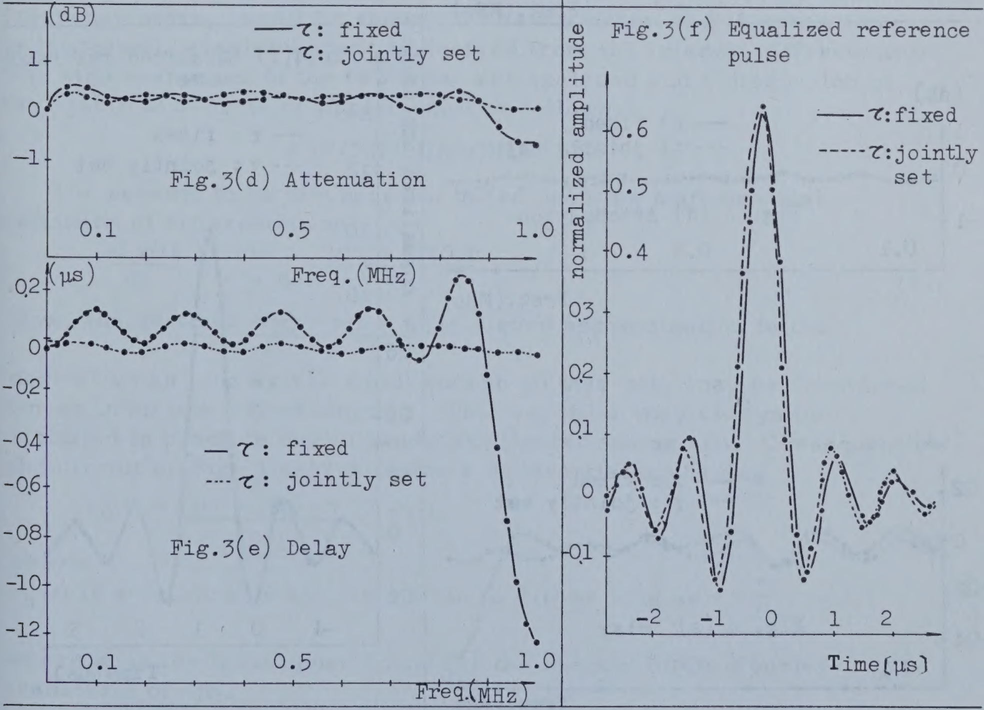


Fig. 3 Simulated and experimental results of a 6dB, 0.56μs slope distortion experiment

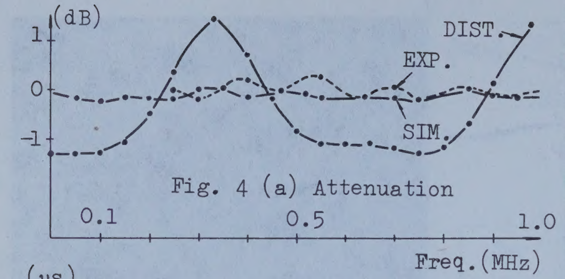


Fig. 4 (a) Attenuation

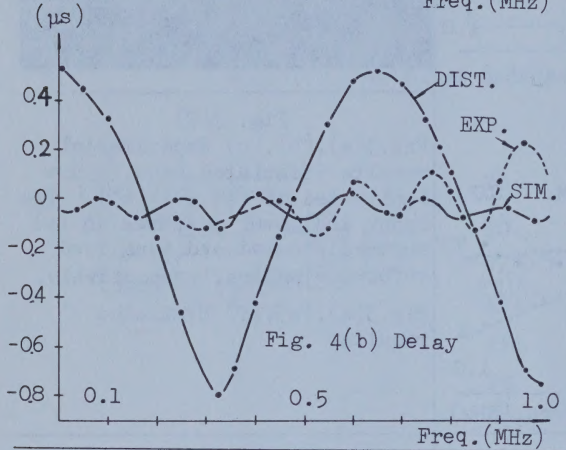


Fig. 4(b) Delay

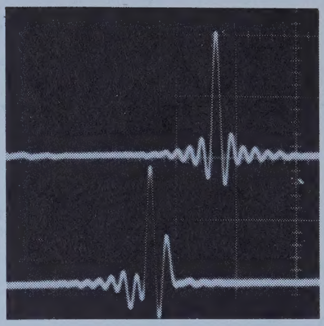


Fig. 4 (C)

Fig.4(a),(b),(c) Experimental results (Simulated results are duplicated in (a) and (b)). The lower and upper pictures in (c) show distorted and equalized reference pulses, respectively

Fig.4(d),(e),(f) Simulated results.

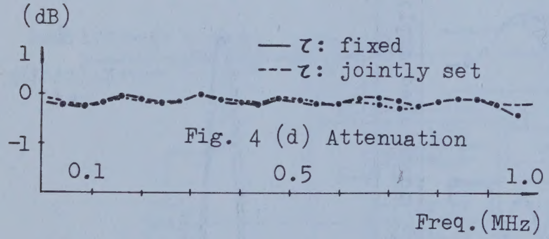


Fig. 4 (d) Attenuation

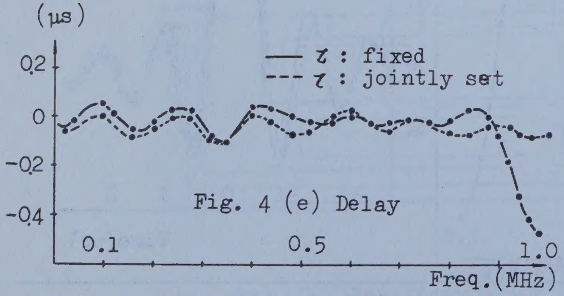


Fig. 4 (e) Delay

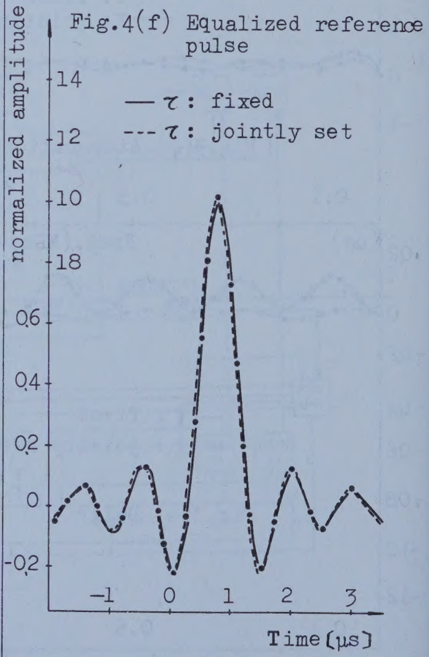


Fig.4(f) Equalized reference pulse

Fig. 4 Simulated and experimental results of a 2.7dB, 1.34μs bump distortion experiment.

AN INFINITE DIMENSIONAL LOW-NOISE DIFFERENTIATOR *

R. D. BROWN AND R. W NEWCOMB

Electrical Engineering Department
University of Maryland
College Park, Maryland 20742

abstract:

The defining limit of the derivative operation provides the basis for the design of a distributed differentiator which is found to have desireable noise characteristics when compared to a lumped RC differentiator. In the presence of white input noise, it is shown that the relative performance of this differentiator improves as the order of differentiation is increased. Two realizations are discussed.

INTRODUCTION

It is well known [1, p. 47] [2, p. 51] that the standard RC differentiators have rather poor noise performance and that this is due to their increased gain with frequency which leads to high frequency noise amplification. Here we introduce delay lines to perform differentiation through the defining equation of the derivative. We will then be exploring the advantages of a differentiator which examines the incoming voltage over a time interval as compared to a standard RC differentiator. Regarding performance in the presence of high frequency noise, it will be shown that as the order of differentiation is increased, greater benefit is derived from the interval differentiator. The step responses of the two types are analyzed and a discussion of the practical aspects of implementation follows.

CIRCUIT DESCRIPTION

The scheme to be discussed is based upon the mathematical definition of differentiation

$$\frac{d\,v(t)}{dt} = \lim_{\tau \to 0} \frac{v(t) - v(t-\tau)}{\tau} \tag{1a}$$

Now, the quotient $\frac{v(t) - v(t-\tau)}{\tau}$ is a good approximation to the derivative as long as τ is small enough so that $v(t)$ may be considered linear in an interval of length τ. This condition may always be satisfied in practice due to bandwidth limitations on $v(t)$. Consequently, the circuit of Fig. 1 approximates a differentiator, having

$$v_o(t) = \frac{v(t) - v(t-\tau)}{\tau} \doteq \dot{v}(t). \tag{1b}$$

where

$v_o(t)$ is a reasonable approximation to $\dot{v}(t)$ as long as $\tau < \frac{\pi}{2\,\omega_c}$

where ω_c is the least upper bound for the support of the Fourier transform of $v(t)$.

---

*This work was supported in part by the Air Force Office of Scientific Research under Grant AFOSR 70-1910.

Second derivatives can be generated from a cascade or alternatively from the circuit of Fig. 2a which is justified through the equations

$$v_1(t) \doteq \dot{v}(t) \tag{2a}$$

$$v_2(t) \doteq \dot{v}(t - \tau) \tag{2b}$$

$$v_0(t) \doteq \frac{v_1(t) - v_2(t)}{\tau} \doteq \frac{\dot{v}(t) - \dot{v}(t - \tau)}{\tau} \doteq \ddot{v}(t) \tag{2c}$$

Higher order derivatives are then found from the differentiator tree of Fig. 2b. Higher order derivatives are generated from information gathered about v(t) over longer time intervals. It is presumed intuitively, and will be demonstrated mathematically that this "time diversity" scheme is beneficial in reducing output noise, compared to cascaded RC differentiators. For such a comparison we will assume that the transfer voltage ratio of the cascade of n RC differentiators is

$$G_{RC}(s) = \frac{V_o}{V_i}(s) = \left(\frac{s}{s + 1/RC}\right)^n \tag{3}$$

FREQUENCY DOMAIN BEHAVIOR

Choosing a delay time of $\tau = 2$ will result in a normalized magnitude response function.

In the sinusoidal steady state we have, for a single normalized delay differentiator,

$$G_1(j\omega) = \tfrac{1}{2}\left(e^{j\omega t} - e^{j\omega(t-2)}\right)/e^{j\omega t} = \tfrac{1}{2}\left(1 - e^{-j2\omega}\right) \tag{4a}$$

or

$$|G_1(j\omega)| = \tfrac{1}{2}\sqrt{(1-\cos 2\omega)^2 + \sin^2 2\omega} = \sqrt{\frac{1 - \cos 2\omega}{2}} \tag{4b}$$

$$\Phi_1(\omega) = \arctan \frac{\sin 2\omega}{1 - \cos 2\omega} = \frac{\pi}{2} - \omega \tag{4c}$$

For comparison purposes we next choose the RC product of the RC differentiator such that $\left.\frac{d|G(j\omega)|}{d\omega}\right|_{\omega=0}$ is the same for it and the delay differentiator.

We find, on using L' Hôpital' s rule,

$$\frac{d\,|G_{RC}(0)|}{d\omega} = RC = \lim_{\substack{\omega > 0 \\ \omega \to 0}} \frac{d\,|G_1(0)|}{d\omega} = 1 \tag{5}$$

Thus, an RC differentiator with low frequency response equivalent to that of a delay differentiator with delay 2 has its corner at $\omega_c = 1$. The responses are shown in Fig. 3 for comparison.

If G_n denotes the voltage transfer function for the n^{th} order tree differentiator we find from direct calculation

$$G_n(s) = [G_1(s)]^n \tag{6a}$$

or

$$|G_n(j\omega)| = |G_1(j\omega)|^n \quad , \quad \Phi_n(j\omega) = n\Phi_1(j\omega) \tag{6b}$$

From these, curves, as for n = 1 of Fig. 3, can be plotted. From such curves one sees that the delay differentiator provides a closer match to the actual derivative for low frequencies while enclosing lesser area at high frequencies.

NOISE PERFORMANCE

In the presence of noise which has a uniform power spectrum, the differentiator tree outperforms cascaded RC differentiators; we next prove a key result in this direction.

For input noise with uniform power spectrum, the noise power output in a frequency band is proportional to the integral of the square of the magnitude voltage transfer function over that band. We may show that for the delay differentiator, this noise power becomes arbitrarily small as the order of differentiation is increased.

Proposition: Given any $\epsilon > 0$, there exists a positive integer N such that for any integer $n > N$, we have,

$$\int_0^\pi |G_n(j\omega)|^2 \, d\omega < \epsilon$$

Proof: Let $\epsilon > 0$ be given; from (4b)

$$\int_0^\pi |G_n(j\omega)|^2 \, d\omega = \int_0^\pi \left(\frac{1 - \cos 2\omega}{2}\right)^n d\omega$$

The integrand achieves a unique maximum of 1 at $\omega = \frac{\pi}{2}$.

Partition the integral as follows:

$$\int_0^\pi |G_n(j\omega)|^2 \, d\omega = \int_0^{\frac{\pi}{2} - \frac{\epsilon}{4}} \left(\frac{1 - \cos 2\omega}{2}\right)^n d\omega \; +$$

$$\int_{\frac{\pi}{2} - \frac{\epsilon}{4}}^{\frac{\pi}{2} + \frac{\epsilon}{4}} \left(\frac{1 - \cos 2\omega}{2}\right)^n d\omega + \int_{\frac{\pi}{2} + \frac{\epsilon}{4}}^{\pi} \frac{(1 - \cos 2\omega)^n}{2} \, d\omega$$

The first and third terms are equal since the integrand is symmetric about $\omega = \frac{\pi}{2}$. In the middle term we may replace the integrand with its maximum value of 1 to arrive at the following inequality:

$$\int_0^\pi |G_n(j\omega)|^2 \, d\omega < 2 \int_0^{\frac{\pi}{2} - \frac{\epsilon}{4}} \left(\frac{1 - \cos 2\omega}{2}\right)^n d\omega + \frac{\epsilon}{2}$$

Now, $\left(\frac{1 - \cos 2\omega}{2}\right)^n$ is monotone increasing in the interval $(0, \frac{\pi}{2} - \frac{\epsilon}{4})$, so we may replace it with its value at the endpoint and continue the inequality.

$$\int_0^{\pi} |G_n(j\omega)|^2 \, d\omega < 2\left(\frac{\pi}{2} - \frac{\epsilon}{4}\right) \frac{(1 - \cos\frac{\epsilon}{2})^n}{2} + \frac{\epsilon}{2}$$

Since $\frac{(1 - \cos\frac{\epsilon}{2})}{2}$ is strictly less than 1, there exists a positive

integer N such that $\frac{(1 - \cos\frac{\epsilon}{2})^n}{2} < \frac{\epsilon/2}{2\left(\frac{\pi}{2} - \frac{\epsilon}{4}\right)}$ for all

$n > N$. So, we have:

$$\int_0^{\pi} |G_n(j\omega)|^2 \, d\omega < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \qquad \text{Q. E. D.}$$

Actually, the N chosen in the above proof is far larger than necessary. The precise value of the integral for any n may be determined from the following expression,

$$\int_0^{\pi} |G_n(j\omega)|^2 \, d\omega = \pi\left(\frac{1}{2}\right)^n \sum_{i=0}^{\substack{n/2 \text{ n even} \\ (n-1)/2 \text{ n odd}}} \left[\frac{n!}{(2i)!\,(n-2i)!}\left(\prod_{k=1}^{i} \frac{2k-1}{2k}\right)\right]$$

At high frequencies, the magnitude response of cascaded RC differentiators assymptotically approaches 1. Thus, most of the high frequency noise is coupled to the output. Now, assuming the noise source generates finite power, there is some frequency ω_c such that 99% of the noise power is generated below ω_c. Then for the tree differentiator, the noise output is proportional to $\int_0^{\omega_c} |G_n(j\omega)|^2 \, d\omega$. This factor diminishes with increasing n. The integrand is periodic in ω and a finite number of periods cover the interval of integration. Thus, by extension of our previous argument, the integral may be made as small as we like by taking higher order derivatives.

The noise performance improves as we increase the size of the tree (in the process of taking higher order derivatives). As can be seen, the area under the $|G_n(j\omega)|$ curve is much less for the tree differentiator than for the RC differentiator. Futhermore, the response of the tree differentiator more closely approximates that of an ideal differentiator. We are always assured of a response peak at $\omega = \frac{\pi}{2}$ for the differentiator tree, whereas corner pileup at $\omega_c = 1$ reduces the effectiveness of cascaded RC differentiators.

PRACTICAL CONSIDERATIONS

1. Magnitude and time scaling

Discussion of the frequency domain behavior of the delay differentiator may have suggested an intimate relationship between delay time and over all level of the magnitude response. To normalize the magnitude peaks we set a delay time of 2 seconds. Of course, the relationship is explicit in the defining equation

$v_0(t) = (v_i(t) - v_i(t - \tau)) / \tau$; τ appears both as delay time and as a scaling factor. After some thought however, one can see that the τ in the denominator has little significance. First of all, the τ as a scaling factor is not required to fill any dimensional needs in the equation. Thus, the transfer function may be scaled up or down simply by changing the basic unit of time. An example will surely clarify this point.

Suppose delay networks are used with delay time of 1 microsecond. Then, measuring time in seconds, we have:

$$v_0(t) = (v_i(t) - v_i(t-10^{-6}))/10^{-6} = 10^{6}(v_i(t)-v_i(t-10^{-6}))$$

Now suppose we measure time in microseconds.

Then, $v_0(t) = v_i(t) - v_i(t-1)$.

Once one accepts the relativity of the scaling factor, it is pleasant to note that if we measure time in blocks of τ , whatever τ may be for the delay network at hand, we get:

$$v_0(t) = v_i(t) - v_i(t-1) .$$

And, for a normalized magnitude function we measure time in blocks of $\frac{1}{2}\tau$, to get: $v_0(t) = (v_i(t) - v_i(t-2))\,\frac{1}{2}$ as previously studied.

This point is vital as a practical consideration. For a delay of 1 microsecond, we are not forced to build a differential amplifier with a voltage gain of a million. We need only change our unit of time.

2. Departures from ideal response

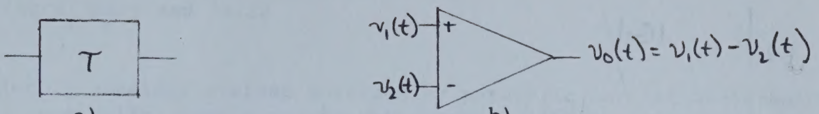
At low frequencies, for a first order differentiator, we would like the output voltage to decrease linearly with decreasing frequency. In theory, at low frequencies the delayed voltage and the initial voltage are nearly equal in magnitude and phase. The difference between the two becomes arbitrarily small as the frequency is reduced. In practice, effective subtraction of nearly equal voltages requires good common mode rejection ratio in the differential amplifier. Also, care must be taken to ensure that the magnitudes of the two voltages are equal, taking into account transmission loss through the delay network, and different sensitivities of the + and - inputs of the differential amplifier. In a differentiator tree these considerations are compounded. At any given level, all of the outputs must have equal magnitude. This requires that the individual amplifiers have good gain stability and that the gain of each one may be trimmed.

In spite of all this, there will still be some frequency below which the output voltage fails to decrease. The desired low frequency rolloff may be continued by capacitor coupling the differential amplifier. This amounts to inserting an RC differentiator with very low corner frequency in such a way that its pole can be made to cancel the effective negative zero of the nonideal delay differentiator, and replace it with a zero at $s = 0$. In a tree with high gain, DC fluctuations caused by power supply or thermal drift get amplified and can cause severe problems in later stages. Capacitor coupling avoids this nicely.

REFERENCES

[1] J. Willman & H. Taub, "Pulse and Digital Circuits," McGraw-Hill, New York, 1956.

[2] R. J. Kochenburger, "Computer Simulation of Dynamic Systems," Prentice-Hall, Englewood Cliffs, 1972.


T
$v_1(t)$
$v_2(t)$
+
-
$v_0(t) = v_1(t) - v_2(t)$

a) Delay

b) Differential Amplifier

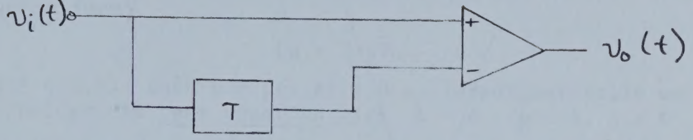
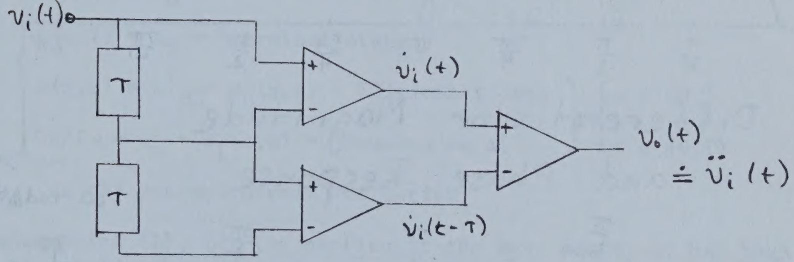


Figure 1
Delay Differentiator


$v_i(t)$
T
T
+
-
$\dot{v}_i(t)$
$\dot{v}_i(t-\tau)$
$v_0(t) \doteq \ddot{v}_i(t)$

a)
Second Derivative

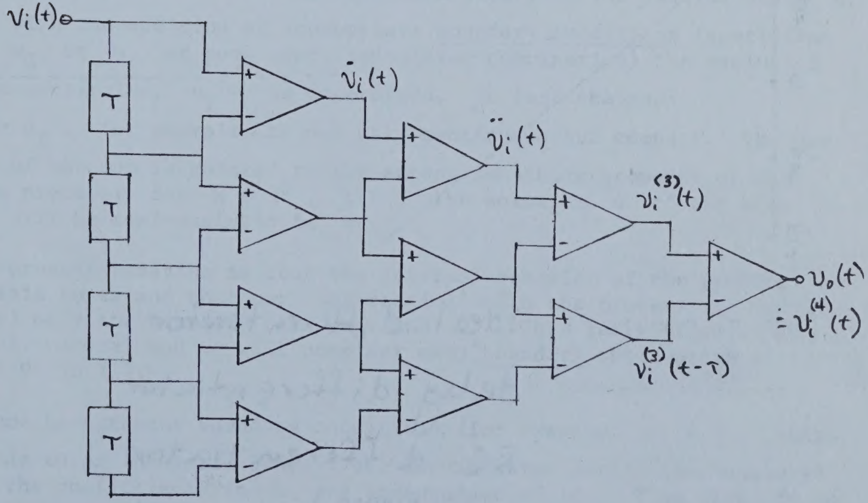

$v_i(t)$
T
T
T
T
+
-
$\dot{v}_i(t)$
$\ddot{v}_i(t)$
$v_i^{(3)}(t)$
$v_i^{(3)}(t-\tau)$
$v_0(t) \doteq v_i^{(4)}(t)$

Figure 2
Differentiator Tree

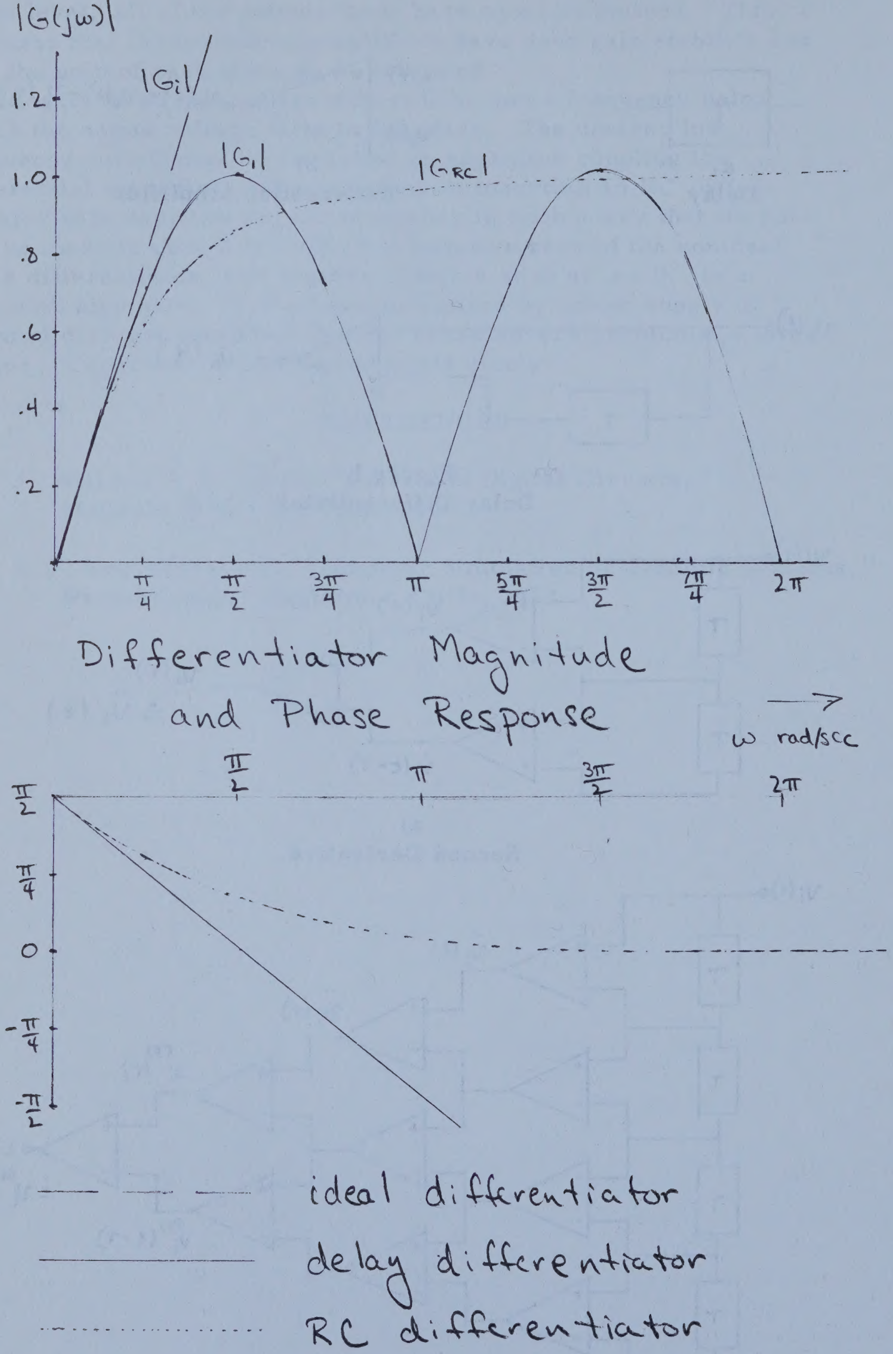


Figure 3
Differentiator Responses

THE 'OBSERVATION AND PREDICTION' PROBLEM FOR DIFFUSION PROCESSES

T. I. SEIDMAN
University of Maryland, Baltimore County
Baltimore, Maryland 21228

We wish to consider systems governed by parabolic partial differential equations; physically, such a system represents a diffusion process. Thus, we have

$$u_t = Lu + f \tag{1}$$

(in which $u = u(x,t)$ for $0 \leqslant t$, $x \in \Omega$ = a bounded, open, connected set in $\mathbb{R}^n$) with L an elliptic operator which we will assume is in 'divergence form'

$$Lu = \nabla\cdot p\nabla u - qu \tag{2}$$

with $q = q(x,t)$ and $p = p(x,t) > 0$. Paradigmatically we may particularly consider the heat equation with $L = \Delta$ ($p \equiv 1$, $q \equiv 0$ in (2)).

We will be considering (1) over a finite time interval $(0,T)$ and will employ the notation:

$$\left\{\begin{array}{l} \left.\begin{array}{l} u(x,0) = u_0 = \text{'initial state'} \\ u(x,T) = u_T = \text{'terminal state'} \end{array}\right\} \quad x \in \Omega \\ \left.\begin{array}{l} u(x,t) = u|_{\partial\Omega} = w_0(x,t) = \text{'Dirichlet data'} \\ \partial u/\partial\nu = u_\nu = w_1(x,t) = \text{'Neumann data'} \end{array}\right\} \left\{\begin{array}{l} x \in \partial\Omega, \\ t \in (0,T) \end{array}\right. \end{array}\right. \tag{3}$$

where u_ν is the outward normal derivative.

The equation (1), (2) (in particular the heat equation) has been the subject of considerable study for some time. Indeed, Fourier invented Fourier series in this connection. It is well-known that (1) does not suffice to determine the evolution of the system from the initial state u_0 but that with the addition of appropriate boundary conditions (specification of w_0 or w_1 or some specified linear combination) the evolution —and, in particular, u_T — is determined. In fact the map: $[u_0,w_0] \mapsto u_T$, for example, is not only continuous but compact. The compactness of the map is related to the strong smoothing property of the diffusion process: for $x \in \Omega^0$, $t > 0$ the solution u of the heat equation will be real-analytic in x.

Our present position is that the internal behavior of the process is inaccessible to us and that our 'interaction' with the process may involve (directly) only the boundary data w_0 and w_1. This restriction seems intuitively natural and we will consider only boundary observation or control with $f \equiv 0$ in (1).

Assume homogeneous boundary conditions (for example: $w_1 \equiv 0$, which corresponds to an insulated body: i.e., no diffusion across the boundary) and that the coefficients in L are independent of t. Then there is an orthonormal basis $\{\phi_k\}$ of eigenfunctions of L with corresponding eigenvalues $\{-\lambda_k\}$ (so $L\phi_k = -\lambda_k\phi_k$). We have, then, the expansion formula

$$u(x,t) = \Sigma_k\ c_k\ \exp[-\lambda_k t]\phi_k(x) \tag{4}$$

in which, of course, the coefficients $\{c_k\}$ must be appropriately determined. This formula (4), together with properties of $\{\phi_k, \lambda_k\}$, has proved central to the study of boundary control and boundary observation problems for (1).

My own introduction to this general topic came by way of a luncheon conversation with Professors MacCamy and Mizel of Carnegie-Mellon University on teaching the notion of a well-posed problem. For the heat equation one may speak glibly in class of taking "an insulated body satisfying the heat equation $u_t = \Delta u$ with 'arbitrary' initial u_0" in order to assert continuous dependence of u_T on u_0 --but Whence cometh this body? How, physically, might one 'create' a body with a desired state?

One is led to consider the following control problem:

> Given a body satisfying $u_t = \Delta u$ with known initial state u_0 and a desired terminal state u_*, can one manipulate the boundary temperature w_0 (alternatively, the boundary flux w_1) for $0 < t < T$ in such a way that $u_T = u_*$? (5)

The real-analyticity of the solution, mentioned above, implies that this is only possible for very special u_* but we were able to show [7], [8] that the reachable states are dense in $L_2(\Omega)$ and even in $C_0^k(\Omega)$, surely adequate for 'practical' purposes. This was shown subject to the additional requirement that the support of the control w_0 was to lie in any specified relatively open subset of $\partial\Omega \times (0,T)$.

If u_0 is not known one might, of course, try to estimate it by observing boundary data for a period of 'natural evolution' in order to follow this by a period of control. One is led to consider the observation problem:

> Given an insulated $(w \equiv 0)$ body satisfying $u_t = \Delta u$ with unknown initial state u_0, can one uniquely determine the state by observation of the boundary temperatures w_0 over $(0,T)$? (6)

It is not difficult to show uniqueness. However, the compactness of the map: $(u_0, w_1 \equiv 0) \mapsto w_0$ guarantees that its inverse cannot be continuous. Thus, if observation of w_0 is only approximate--and any realistic discussion of observability must include consideration of 'errors of measurement'--it is impossible to bound the resulting uncertainty in the estimation of u_0 ; no such estimation procedure can be 'stable'. Fortunately, the same compactness (smoothing) property which makes estimation of u_0 unstable, affords the possibility that one can stably estimate u_T, i.e., predict the terminal state. This is the *observation and prediction problem* which we wish to consider and we interpret (6) to mean consideration of the map: $w_0 \mapsto u_T$ with $w_1 \equiv 0$ and u_0 unknown.

It is interesting, in the context of our present symposium, to observe that the difficulty in the estimation of u_0 is intrinsically 'an infinite-dimensional consideration.' The uniqueness noted above means existence of an inverse which, if the state space were finite-dimensional, would necessarily be continuous. Only in the infinite-dimensional case does a problem remain at this point in the absence of, e.g., stochastic perturbation of the governing differential equation.

Similar considerations apply to any *compact process*--i.e., any linear system with infinite dimensional state space for which the evolution map $S_{t,\tau}$ (carrying an initial state at time τ into a terminal state at time t) is a compact linear operator for $t > \tau$. Examples include not only diffusion processes but also systems governed by delay-differential equations.

The problem (6) of boundary observation and prediction leads, through the basic expansion formula (4) to concern with questions regarding functionals of Dirichlet series. It follows from (4) that

$$w_0(t) = \Sigma_k\, c_k\, \bar{\phi}_k \exp[-\lambda_k t] \ , \qquad t \in (0,T) \tag{7}$$

where each $\bar{\phi}_k$ is the 'trace of ϕ_k' on $\partial\Omega$ -- i.e., the $\{\bar{\phi}_k\}$ are the Dirichlet data for the Neumann problem ($w_1 \equiv 0$) eigenfunctions of L. Note that the domain of the operator $A : w_0 \mapsto u_T$ under consideration in (6) is not all of $L_2(\partial\Omega \times (0,T))$ but is, rather, the subspace M of functions having Dirichlet expansions such as (7). (These are quite special functions: e.g., each $w \in M$ extends to a function analytic on $\mathrm{Re}\, t > 0$.) Since u_T has the orthogonal expansion

$$u_T(x) = \Sigma_k\, c_k \exp[-\lambda_k T]\phi_k(x) \ , \qquad x \in \Omega \ , \tag{8}$$

we see that A can be constructed from the coefficient functionals

$$\psi_j : w_0 \mapsto c_j \tag{9}$$

defined on M. By the Hahn-Banach Theorem, such functionals are well-defined and continuous provided the functions $\{\bar{\phi}_k \exp[-\lambda_k t]\}$ are 'free' in $\mathcal{L}_2(\partial\Omega \times (0,T))$--indeed, the norm of each functional ψ_j is just the reciprocal of the distance d_j from the function $\bar{\phi}_j \exp[-\lambda_j t]$ to the span M_j of the *other* such functions.

If we restrict our observation further, from all of $\partial\Omega$ to a single point ξ (e.g., in the one-dimensional case we might have $\Omega = (0,1)$ and $\xi = 0$), $\phi_k = \phi_k(\xi)$ in (7) is just a scalar. By the Müntz-Szász Theorem [12], [18], coefficient determination in a Dirichlet series is possible if and only if

$$\Sigma_k\, 1/\lambda_k < \infty \ . \tag{10}$$

In the *one-dimensional case* one has [2] asymptotically

$$\lambda_k = c^2k^2 + \mathcal{O}(k) \tag{11}$$

so (10) holds. The Müntz-Szász Theorem would apply (and give estimates $\{N_j\}$ for the norms of the coefficient functionals) only for $T = \infty$. Fortunately, a result of L. Schwartz' [15] permits use of this for $T < \infty$ with the estimate multiplied by K_T (depending on T but not on j). Substituting in (8) gives

$$\|Aw_0\|^2 \leqslant \Sigma_j (N_j/|\bar{\phi}_j|)^2 \exp[-2\lambda_j T]\,\|w_0\|^2 \tag{12}$$

and (as, from (11), $N_j = \exp[\mathcal{O}(j)]$) this bounds $\|A\|$ --unless $\bar{\phi}_j$ were to go (rapidly) to 0. For the one-dimensional heat equation on $\Omega = (0,1)$ (cf., [10]) one has $\phi_0 = 1$, $\phi_k(x) = (\cos k\pi x)/\sqrt{2}$ so $|\bar{\phi}_k| = \phi_k(0) \geqslant 1/\sqrt{2}$ and (12) bounds A. For the one-dimensional case with the more general operator (2) one still has $\{\phi_k(0)\}$ bounded away from 0 (c.f., [16]) and again A is bounded by (12).

For the n-dimensional case $(n > 1)$ the asymptotic estimates for λ_k show that (10) fails. A modification of the argument above can still be used under special circumstances. Suppose there were an orthonormal sequence $\{\theta_j\}$ in $L_2(\partial\Omega)$ and a partition of the integers into sequences Z_j such that

$$\begin{aligned}
&\text{(a)}\quad \left\langle \bar{\phi}_k, \theta_j \right\rangle_{\partial\Omega} = 0 \quad \text{if} \quad k \notin Z_j\,,\\
&\text{(b)}\quad |\gamma_k| \geqslant c > 0 \quad \text{with} \quad \gamma_k = \left\langle \bar{\phi}_k, \theta_j \right\rangle_{\partial\Omega} \quad \text{for} \quad k \in Z_j\,,
\end{aligned} \tag{13}$$

(c) for each sequence $\Lambda_j = \{\lambda_k : k \in Z_j\}$, the original argument applies.

The scalar functions $W_j(t) = \left\langle w_0(t), \theta_j \right\rangle_{\partial\Omega}$ take the form

$$W_j(t) = \Sigma\{c_k\, \gamma_k \exp[-\lambda_k t] : k \in Z_j\}$$

so we have bounded operators $A_j : W_j \mapsto U_j \in L_2(\Omega)$ with $U_j = \Sigma\{c_k \exp[-\lambda_k T]\phi_k : k \in Z_j\}$. Thus

$$\|Aw_0\|^2 = \Sigma_j \|U_j\|^2 \leqslant \Sigma_j \|A_j\|^2 \|W_j\|^2 \tag{14}$$

$$\text{as}\quad \Sigma_j \|W_j\|^2 \leqslant \|w_0\|^2$$

$$\leqslant (\sup_j \|A_j\|)^2 \|w_0\|^2 .$$

For the heat equation in a cylinder $(\Omega = (0,1) \times \mathcal{D})$ [10], all the Λ_j are the same and (14) bounds A. In the case of the n-ball, however, the Λ_j differ and it was necessary in [11] to show that $\{N_j\}$, K_T -- which, of course, depend on the exponent sequence--can be estimated uniformly for the class of sequences $\{\Lambda_j\}$. (The paper [5] by Fattorini and Russell improves this result to give larger classes of sequences in connection with a boundary *control* problem as in (5).)

The method based on (13) depends quite strongly, for its applicability, on the separability of Δ in special geometries. In general, no such sequence $\{\theta_j\}$ will exist. While the validity of the *result* intuitively should not depend on such separability, the technique of proof clearly does.

To have any chance of extending the basic approach via (7), (8) would require a deeper understanding than is now available of H-valued Dirichlet series (H a hilbert space, as $\mathcal{L}_2(\partial\Omega)$) . Estimates are needed for functionals involving coefficients of series $\Sigma_k c_k \exp[-\lambda_k t]\xi_k$ $(\xi_k \in H)$ for appropriate sequences $\{(\lambda_k,\xi_k)\}$. This program, for application to (6), would also require investigation of properties, as elements of $\mathcal{L}_2(\partial\Omega)$, of Dirichlet data for Neumann problem eigenfunctions.

Problems of both boundary control and boundary observation/prediction lead, through (7), to concern with quite similar questions regarding Dirichlet series. This similarity might be anticipated in view of the 'duality' known for observability and controllability in linear system theory. Here we can give explicit content to this heuristic duality (for convenience, stated here only in a restricted form).

Given the (possibly time-dependent) elliptic operator L of (2), define L' by

$$L'v = \nabla \cdot \hat{p}\nabla v - \hat{q}v \tag{15}$$

with $\hat{p}(x,t) = p(x,T - t)$, $\hat{q}(x,t) = q(x,T - t)$ and consider the Neumann control problem of (5) with terminal state 0 . For any $v_0 \in \mathcal{L}(\Omega)$ there is (cf., [13]) a control ω , in general not unique, so that v satisfying $v(\cdot,0) = v_0$, $\partial v/\partial\nu = \omega$, and $v_t = L'v$ also has $v(\cdot,T) \equiv 0$; now set $\hat{\omega}(x,t) = -\hat{\omega}(x,T - t)/p(x,t)$. We have the following duality theorem [17].

Theorem: The linear functional (16)

$$\psi_0 : w \mapsto \int_0^T \int_{\partial\Omega} \hat{\omega}(x,t)w(x,t)dx\,dt \ , \quad w \in M$$

is independent of the non-uniqueness of ω . The map: $v_0 \mapsto \psi_0$ is just A^* , the *adjoint* of the map A of (6). Since $||A^*|| = ||A||$, a consequence of this theorem is that any suitable estimate for $||\omega||$ in terms of $||v_0||$ would automatically bound A and show the well-posedness of the observation/prediction problem. Indeed, based on [8], [10], [13], [14] one might anticipate well-posedness of (6) with observation restricted to any relatively open subset of $\partial\Omega \times (0,T)$.

A truly impressive recent paper by Russell [13] introduces a novel approach to this program for $L = \Delta$. By the method of spherical means, results are obtained for boundary controllability of a process governed by the *wave equation* $u_{tt} = \Delta u$ in Ω . This is related to 'harmonic analysis' for expansions of the form: $\Sigma_k c_k \exp[\pm i\omega_k t]\bar{\phi}_k$ $(\omega_k^2 = \lambda_k)$. A Fourier transform technique obtains, from this, Dirichlet series results of the very sort we need, which are then applied to derive boundary controllability results for the heat equation. This approach is used in [13] to obtain only existence, with no estimates except for quite special v_0, of the control ω. One must hope that careful consideration of the approach would give the needed estimates more generally. One interesting application of the Theorem (16) to the boundary control problem is that among the controls ω taking v_0 to 0 there is one (in fact, the one of minimal $L_2(\partial\Omega \times (0,T))$ norm) in M, at least for $L = \Delta$.

We have, so far, considered only the problems (5) and (6) and their relation. There are, however, numerous interesting variations. A rather exotic possibility, for example, would be to observe, over (0,T), the motion of the free boundary in a Stefan problem and so determine the state at T.

Closely related to (6) would be prediction of u_T given observation of a (consistent) pair (w_0, w_1) with $w_1 \neq 0$. Without a fixed homogeneous boundary condition there is no expansion formula (4) but we may attempt the following approach. Let $\hat{u}$ satisfy (1) with 0 initial data and Neumann data w_1; then $\hat{u}$ will have some Dirichlet data $\hat{w}_0$. This reduces the more general problem to that already considered as $(w_0 - \hat{w}_0)$ may now be used to predict the terminal state of $(u - \hat{u})$, which has 0 Neumann data, and then $\hat{u}_T$ added. For the reduction to be valid, however, would require a guarantee that the map: $w_1 \mapsto \hat{w}_0$ is continuous. I would, indeed, conjecture that this map is a homeomorphism but this is still a conjecture.

Another interesting variant is prediction on the basis of internal observation (cf., [1], [3]) At least in the one-dimensional case, this is even physically reasonable. Given observation of $w_\xi = u(\xi,t)$ for $\xi \in \Omega = (0,1)$, one still has (7) with $\bar{\phi}_k = \phi_k(\xi)$. For ξ a zero of any ϕ_k the problem is clearly impossible and even for other ξ one might still have $\phi_k(\xi) \to 0$ rapidly. For the heat equation, Dolecki [3] has analyzed the situation to show that A is bounded for $\xi \notin S = S(T)$ with S of measure 0 but dense (it includes the rationals) and determined by complicated number-theoretic criteria; observation by a moving observer $(w_*(t) = u(\xi_0 + ct,t))$ is well-posed for any T.

One might also consider [16] observation of a local average:

$$w_\theta(t) = \int_0^1 u(x,t)\theta(x)dx , \qquad t \in (0,1) . \tag{17}$$

Formula (7) still applies, with $\bar{\phi}_k = \int \phi_k\theta$ and we ask whether $\bar{\phi}_k = 0$ or $\bar{\phi}_k \to 0$. For θ a step-function or chapeau-function centered at ξ with half-width α one has the same sort of considerations as Dolecki obtained for point observation--but now applying both to ξ and α. Even more curious is the situation for θ a modified (for ease of computation, due to the use of the finite interval Ω) Gaussian distribution centered at ξ with standard deviation α. In this case one has exactly Dolecki's restrictions on ξ but now must also impose the condition: $T > \alpha^2/2\pi^2$ for A to be bounded.

Returning to the original problem (6), observe that we know $A = A_T : w_0 \mapsto u_T$ is bounded in, e.g., the one-dimensional case, but obviously the bound must depend on T. Clearly $||A_T|| \to \infty$ as $T \to 0$; but how rapidly? E.g., does $T||A_T|| \to 0$. This would be significant for any attempt at numerical computation based on subdivision of the time interval. Similarly, estimates would be useful--in connection, say, with a problem involving time-dependent coefficients, for which (4) is not directly available--of the uncertainty in u_T induced by alteration of the sequence of pairs $\{(\phi_k, \lambda_k)\}$. Finally, we remark that while we have here presented results only in an L_2 setting, the problems can all be considered--and many have been--using more general norms.

REFERENCES

[1] J. R. Cannon and G. H. Knightly, "The approximation of the solution to the heat equation in a half-strip from data specified on the bounding characteristics," SIAM J. Numer, Anal., 6 (1969), 149-159.

[2] R. Courant and D. Hilbert, "Methods of Mathematical Physics, vol. I," Interscience, N.Y., 1953.

[3] S. Dolecki, "Observation for the one-dimensional heat equation," Inst. Math. (Polish Acad. Sci.), preprint 42, 1972.

[4] H. O. Fattorini and D. L. Russell, "Exact controllability theorems for linear parabolic equations in one space dimension," Archive Rat. Mech. and Anal., 43 (1971), 272-292.

[5] __________, "Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations," Q. of Appl. Math., to appear.

[6] S. Kaczmarz and H. Steinhaus, "Theorie der Orthogonalreihen," Mono. Mat. (tom VI), Warsaw, 1935.

[7] W. A. J. Luxemburg and J. Korevaar, "Entire functions and Müntz-Szász type approximation," Trans. A.M.S., 156 (1971), 23-37.

[8] R. MacCamy, V. J. Mizel and T. I. Seidman, "Approximate boundary controllability for the heat equation," J. Math. Anal. Appl., 23 (1968), 699-703.

[9] __________, "Approximate boundary controllability of the heat equation, II," J. Math Anal. Appl., 28 (1969), 482-492.

[10] V. J. Mizel and T. I. Seidman, "Observation and prediction for the heat equation," J. Math Anal. Appl., 28 (1969), 303-312.

[11] __________, "Observation and prediction for the heat equation, II," J. Math. Anal. Appl., 38 (1972), 149-166.

[12] C. Müntz, "Über den Approximationenssatz von Weierstrass," Math. Abh. (Sch. Festschr.), (1914), 303-312.

[13] D. L. Russell, "A unified boundary controllability theory for hyperbolic and parabolic partial differential equations," Studies in Appl. Math., to appear.

[14] ________, "Exact boundary value controllability theorems for wave and heat processes in star-complemented regions," to appear.

[15] L. Schwartz, "Étude des sommes d'exponentiélles, 2-me édition," Hermann, Paris, 1959.

[16] T. I. Seidman, "Observation and prediction for one-dimensional diffusion equations," to appear.

[17] ________, "A duality theorem for boundary observation and control of diffusion processes," to appear.

[18] O. Szász, "Über die Approximation stetiger Funktionen durch lineare Aggregate von Potenzen," Math. Ann., 77 (1915), 482-496.

OPERATOR THEORY TECHNIQUES FOR DISTRIBUTED SYSTEMS

J. WILLIAM HELTON
Department of Mathematics, S.U.N.Y.
Stony Brook, New York 11790

Systems realizability theory, a branch of scattering theory, and the canonical models branch of operator theory developed independently of each other, and later were seen to be equivalent. The connection between Lax-Phillips scattering theory and canonical models was discovered about ten years ago. That systems theory belongs in this circle was discovered more recently. Note that the Russian operator theorist, M. S. Livsic, initiated a treatment of systems in the early sixties which seems to be independent of the American development of the subject [12].

This talk will outline a scattered collection of topics in this area which reflect my particular interests. The first section discusses infinite dimensional realizability, the second connections with the Nagy-Foias model theory, the third describes how one can treat systems (linear or non-linear) from a scattering theory viewpoint. For example, one can associate scattering operators, etc., with a programmed computer. The fourth observes that one can treat the distributed optimal regulator with the realizability theory of §1 along with the rather powerful spectral factorization theorem worked out by operator theorists. The last section gives some miscellaneous remarks and opinions. We shall always work with discrete time.

I would like to thank Mrs. Virginia LaLumia for typing this manuscript.

§1. Realizability Theory

Now the conventions for the talk. We shall work with a discrete time system [A,B,C,D] having state space X, input space U, and output space Y. That is,

$$x(n+1) = A\,x(n) + B\,u(n)$$

$$y(n) = C\,x(n) + D\,u(n)$$

The frequency response function is

$$T(z) = D + zC(I-zA)^{-1}B$$

which is assumed to be defined for z in a neighborhood of the origin. The controllability operator is a map

$$\ell^2(0,\infty,U) \rightarrow X \quad \text{defined by}$$

$$C(u_o,u_1,u_2,\ldots) = \sum_{k=0}^{\infty} A^k B u_k$$

The system [A,B] is called approximately (exactly) controllable if the range of C is dense in X (is all of X). Exact controllability is equivalent to the standard pseudo-inverse C^{-1} of C being a continuous operator. One can also set the same definitions for the adjoint system to [A,B,C,D] and thereby obtain the observability operator Q and notions of approximate and exact observability.

Now let us recall the high points of discrete time realizability theory for infinite dimensional systems (see [1], [2], [3], [5], [7], for various treatments of the general realizability theory). The following theorem treats uniqueness of a realization for a particular frequency response function.

Uniqueness Theorem. If $[A,B,C,D]$ and $[\dot{A},\dot{B},\dot{C},\dot{D}]$ are two approximately controllable and observable systems with the same frequency response function, then there exists a one-one operator M with dense range such that

$$MA = \dot{A}M \qquad MB = \dot{B} \qquad \dot{C}M = C \qquad D = \dot{D}$$

One can take the operator M to be

$$M = (\dot{Q}^{-1})^*Q^* = \dot{C}C^{-1} .$$

Clearly, various assumptions on controllability and observability give various properties of M. For example, if both systems are exactly controllable or observable, then M is bounded and invertible.

Among operator theorists it is well-known that when M is bounded and invertible virtually all properties of A and $\dot{A}$ must be the same, whereas, if M is merely one-one with dense range then, if the domain of M behaves well, the eigenvalues of A and $\dot{A}$ must be the same while absolutely everything else can change. Intermediate cases vary in between the extremes usually on the side of preserving little structure (see, for example, page 70 and Ch. IX [13]).

Realizability Theorem. Any operator valued function $F(z)$ analytic and bounded on the unit disk is the frequency response function for some exactly observable (controllable) and approximately controllable (observable) system. It has a realization which also is exactly controllable (observable) if and only if Hankel_F has closed range.

One can actually write down the systems accomplishing the realization. Let $L^2(H)$ denote the Hilbert space H valued functions on the unit circle with square integrable norms and let $H^2(H)$ (resp. $\bar{H}^2(H)$) denote the subspace of functions with zero negative (resp. non-negative) Fourier coefficients. Let M_F denote the operator $L^2(H) \rightarrow L^2(H)$ defined by $[M_F g](e^{i\theta}) = F(e^{i\theta})g(e^{i\theta})$ for all $g \in L^2(H)$ and let P_S denote the orthogonal projection of $L^2(H)$ onto the subspace S of $L^2(H)$. Frequently, we will write $L^2, H^2, \bar{H}^2$ instead of $L^2(H)$, etc. An exactly observable realization of F is

A is the restriction of $P_{H^2} M_{e^{-i\theta}}$ to the subspace

$X = P_{H^2} M_F \bar{H}^2$

$B : H \rightarrow X$ is given by $Bh = P_{H^2} M_{F(e^{i\theta})} h e^{-i\theta}$

C is the projection of X onto the subspace of constant functions in H^2.

D is $F(0)$.

The reader might note that there is a nice criterion for determining which matrix valued functions can be realized by exactly controllable and observable systems. It was stated originally for scalar functions by Douglas Clark and myself and then extended to the matrix case and to continuous time systems by Paul Fuhrmann. The interested reader should see Fuhrmann's [6] or his talk appearing in this volume.

§2. The Nagy-Foias Model Theory

The Nagy-Foias Model Theory is to a systems theorist a study of a special class of discrete time systems. In their notation the state space for the system is H, the input space is $\mathfrak{D}_{A^*}$, the output space $\mathfrak{D}_A$ all Hilbert spaces infinite or finite dimension. The systems $[A,B,C,D]$ they study satisfy the special relationships

$$\|A\| \leq 1, \quad B = (1-AA^*)^{1/2}, \quad C = (1-A^*A)^{1/2}, \quad D = A^*|_{\mathfrak{D}_{A^*}} .$$

These were dubbed standard unitary (s.u.) systems by the author. One can show that a s.u. system is lossless if and only if it is exactly controllable and observable and any lossless system is a s.u. system up to changes allowed by the Uniqueness Theorem. The first seven chapters of the book by Nagy-Foias [13] are devoted to general properties of such systems. For example, Theorem VI 3.1, [13] is the realizability theorem while Theorem VI 3.4, [13] is the uniqueness theorem for such systems. For lossless systems, these theorems follow from the realizability and uniqueness above. The last half of Chapter 7, Chapter 8, and the first half of Chapter 9 in [13] are less general, and they apply directly to systems with finite dimensional input and output space. The last half of Chapter 9 pertains to transforming continuous time to discrete time systems.

The objective of [13] is, of course, not to do systems theory. The objective is to determine when two operators A_1 and A_2 on a Hilbert space are unitarily equivalent, that is, $MA_1M^{-1} = A_2$ for M a unitary operator. The answer Nagy-Foias obtain is that A_1 and A_2 are equivalent if and only if the s.u. systems correspond to them having the same frequency response functions. The Realizability and Uniqueness Theorems give, in my opinion, the correct canonical model similarity (quasi-similarity, etc.) theory at least for operators A with A and A* both assymptotically stable. Evidence for this has been given recently by Fuhrmann and Moore, and also by myself, §4 in [7].

§3. Scattering Theory

A typical situation in scattering theory is that you wish to determine the constitution of an obstacle at some distance from you by bouncing sound off of it. This is called the inverse scattering problem. What you do is compare the behavior of sound if no obstacle is present to the behavior of sound if an obstacle is present. There are certain objects which arise from this comparison and which are always associated with a scattering problem: the forward and backward wave operators and the scattering operator.

Abstract scattering theory is very closely related to systems theory. Given a particular obstacle one can show that there is a distributed parameter linear system associated with it (see [7] for a discussion) which contains precisely the information contained in the scattering problem. Thus, a scattering problem can be viewed in a conventional systems theoretic way. The reverse of this is also true: systems theory can be treated in a scattering theoretic way. Recently, my wife and I proved that one could treat non-linear systems in a scattering theoretic way and that, for example, it is quite natural to associate wave and scattering operators with a programmed computer [7]. That paper includes a brief exposition of scattering theory.

The formal structure is as follows. To simplify exposition we make the inessential assumption that $U = Y \subset X$. Consider a non-linear system,

$$x_{i+1} = \delta(x_i,u_i) \quad \text{and} \quad y_i = \lambda(x_i,u_i)$$

with the additional property that there is a letter - such that $- = \delta(-,-)$ and $- = \lambda(-,-)$. Most of the action takes place on the universe Ω' of all (input strings, states, output strings) for a system; a typical element w of Ω'

$$(\ldots,u_{-2},u_{-1},\boxed{x},y_1,y_2,\ldots)$$

Here a □ is always drawn to distinguish the element in the state space. The system acts on Ω' in the obvious way, namely, it maps w to

$$(\ldots,u_{-2},\boxed{\delta(x,u_{-1})},\lambda(x,u_{-1}),y_1,y_2,\ldots)$$

and the map of Ω' to Ω' which implements this, called the time evolution operator, is denoted by W. One example is a free system. For such a system $X = Y = U$ and the evolution operator for it is the shift S_+ by one unit to the right. In acoustic scattering this structure occurs when no obstacle is present; for computers a free system corresponds to a reset machine (slightly modified flip-flop).

One now compares the original system to a free system. For convenience let $\Omega = \{w \in \Omega' :$ state component of w is in $U\}$, let S_+ be the shift on Ω, and regard this as the desired free system. The backward wave operator is defined to be

$$W_- = \lim_{n\to\infty} W^n S_-^n$$

A similarly constructed arrangement is used to define the formal inverse of the forward wave operator

$$W_+' = \lim_{n\to\infty} S_-^n W^n$$

and the scattering operator is $\Sigma = W_+ W_- S_-$. These limits exist on large classes of sequences in Ω and Ω' respectively and they can be computed explicitly. An amusing exercise is to write down the scattering operator for simple devices such as and-or gates, pulse adders, and comparators.

These objects are canonical and could be used to study the system. For linear systems Σ is actually just a transform of the frequency response function. Moreover, W_+' and W_- carry the same type of information as the controllability and observability operator, and they can be expressed simply in terms of them.

One can also show that the scattering operator has the expected properties of time invariance and causality where they are suitably defined. It is possible to prove (using the Nerode equivalence technique) that any time invariant causal map on a suitable universe Ω can be realized as the scattering operator of some time invariant system.

Off hand it seems to me that one could study systems theory using either the scattering or conventional approach and the fact that it developed one way as opposed to the other may possibly just be a historical accident. Unifications of this sort probably have fairly immediate pedagogical interest and hopefully some techniques and examples treated by scattering theorists will prove useful to systems theorists. One effect of systems theory on mathematical scattering theory will probably be to broaden its horizons considerably. Classical scattering theory is devoted entirely to lossless systems and only recently (cf. [11]) has a theory for lossy systems emerged. My limited knowledge of systems theory, its myriad examples and very general theory, make me suspect that scattering theory actually has a broader scope than mathematicians have realized.

§4. A Spectral Factorization Approach to Distributed Control

We consider the system

$$x_{i+1} = Ax_i + Bu_i \qquad (4.1)$$

where A is stable, that is, $\|A^n\| \leq M$ and assume that the cost of running the system with control $\langle u_i \rangle_1^N$ from state

x_o for N time units is

$$J_N(x_o,u) = \sum_{i=0}^{N} [(x_i,Qx_i) + (u_i,Ru_i)] \qquad (4.2)$$

Here $Q : X \to X$ and $R : U \to U$ are bounded self-adjoint operators. The optimal regulator problem is given a state x_o determine a control sequence $u \in \ell^2(0,\infty,U)$ which minimizes the cost $J_\infty(x_o,\cdot)$ and for which $x_n \to 0$ in some sense as $n \to \infty$.

The optimizing rule for such a problem is frequently expressed in terms of a feedback law, namely, $u_i = \Gamma x_i$ for some operator $\Gamma : X \to U$. A presentation of the following material which is more complete than this sketchy and sometimes ambiguous one will appear in [9].

Sometimes it is convenient to regard our system as having an output. The natural choice for the output operators is $|Q|^{1/2}$ and our control problem is equivalent to minimizing the cost $\Sigma(y_i,\text{sgn}Qy_i) + (u_i,Ru_i)$ of running the system $x_{i+1} = Ax_i + Bu_i$ and $y_i = |Q|^{1/2}x_i$. The frequency response function for the system is $W(z) = z|Q|^{1/2}(I-zA)^{-1}B$, which is defined and analytic inside the disk. We shall assume henceforth that it is uniformly bounded. Thus, radial limits (see Ch. V, §2, [ ]) of W exist a.e. to give a function $W(e^{i\theta})$ defined on the unit circle.

Now some definitions. A trajectory of the system [A,B] is a X-valued sequence $\{x(n)\}_{n=0}^{\infty}$ which results from feeding some $\ell^2(0,\infty,U)$ input sequence into the system (4.1). Let $\tilde{\mathcal{R}}$ denote the image of all finite length sequences in $\ell^2[0,\infty,U)$ under the controllability operator. A uniformly bounded function for the circle with analytic extension onto the disk is called outer if range $M_F|_{H^2}$ is dense in H^2.

Let $u(e^{i\theta})$, $x(e^{i\theta})$, etc. denote the Fourier transform of sequences $\{u_n\}$, $\{x_n\}$, etc. and notice that Fourier transforming (4.2) gives $J_\infty(0,u) = \int_{-\pi}^{\pi}(u(e^{i\theta}),E(e^{i\theta})u(e^{i\theta}))d\theta$ where

$$E(e^{i\theta}) = R + W(e^{i\theta})^*\text{sgn}QW(e^{i\theta})$$

Thus, E gives the cost of operating the system with initial state zero. The statement $E(e^{i\theta}) \geq 0$ roughly means that the zero state stores no energy, and it is reasonably close to saying that E has a spectral factorization. Here spectral factorization means $E(e^{i\theta}) = M(e^{i\theta})^*M(e^{i\theta})$ where $M(z)$

is analytic in the disk and uniformly bounded. Exactly which non-negative functions have spectral factorizations is a classical question. Work sometimes cited in engineering literature is that of Gokberg-Krein in the mid-fifties. Actually necessary and sufficient conditions are available (see [14], and Ch.V §4 [13]). One sufficient condition is that E have a meromorphic pseudo-continuation onto all of $\mathbb{C}$ [4]. Anyhow, there are quite a few ways to check if a function E has a spectral factorization and so the hypothesis of the following theorem is hopefully practical to apply.

Theorem 4.1. If E has a spectral factorization, then there exists a feedback law $u = \Gamma x$ with Γ defined on range C, which gives the optimal performance of the system. Moreover, there exists a bilinear form K(,) defined on $\tilde{\mathfrak{R}}$ so that K(x,x) gives the cost of running the system initially at $x \in \tilde{\mathfrak{R}}$ forever, with the optimal control law. The proof is in principle constructive. Also, along any trajectories of the feedback system $K(x_N,x_N) \to 0$.

We shall call the optimal cost functional (o.c.f.) The main idea behind the proof goes as follows: If E has a spectral factorization M*M, where M has a realization of the form $G + zF(I-zA)^{-1}B$, then one can show that $\Gamma = - G^{-1}F$ gives the desired control law and that

$$K(x,y) = \sum_{j=0}^{\infty} (A^j x,[Q-F^*F]A^j y)$$

when properly interpreted is the o.c.f. Spectral factorizations of E are not unique and some of them have no such realization, however, one can prove that a factorization M(z) which is an outer function, has such a realization. This is accomplished with the construction behind the Realizability Theorem (suitably applied). The key is to keep careful track of the state spaces in the restricted shift realization for $\tilde{W} = W(e^{-i\theta})^*$ and $\tilde{M} = M(e^{-i\theta})^*$. Since M is outer $\zeta_1 =$ clos $(\mathfrak{m}_{\tilde{M}}\bar{H}^2) \cap H^2$ equals clos $(\mathfrak{m}_{E(e^{-i\theta})}\bar{H}^2) \cap H^2$, which in turn is contained in $\zeta_2 =$ clos $(\mathfrak{m}_{\tilde{W}}\bar{H}^2) \cap H^2$; these are the appropriate state spaces. The fact that $\zeta_1 \subset \zeta_2$ along with an application of the Uniqueness Theorem to $\tilde{W}$ yields that M has the desired relation.

The stability analysis of the resulting feedback system rests on the identities

$$(I-z[A+B\Gamma])^{-1}BG^{-1}M(z) = (I-zA)^{-1}B \qquad (4.3)$$

$$M(z)G^{-1}F(I-z[A+B\Gamma])^{-1} = F(I-zA)^{-1} \qquad (4.4)$$

and from them one can obtain

Theorem 4.2. Suppose that $E(e^{i\theta}) \geq \delta I$ where $\delta > 0$. Then the trajectories of $[A,B]$ are precisely equal to the trajectories of the feedback system $[A+B\Gamma,B]$ if and only if M is an outer function. Furthermore, if M is outer, the range of the controllability operators for these two systems are equal. A similar statement holds for observability of $[F,A+B\Gamma]$.

Thus, if all trajectories of the original system tend to zero, then all trajectories of the feedback system do also. If the original system is exactly controllable, then so is the feedback system. The last statement of the theorem, along with the construction in Theorem 4.1. give $[F,A+B\Gamma]$ is exactly observable if and only if $P_{\bar{H}^2}\mathfrak{m}_E P_{H^2}$ has closed range. These controllability and observability statements imply that $(A+B\Gamma)^*$ and $A+B\Gamma$ respectively, are assymptotically stable, (see [5], or §4 of [8]).

§5. Miscellaneous Comments

A complete mathematical theory should be composed of several layers. There should be general theorems giving the basic structure common to many situations and then there should be a branched hierarchy of increasingly special conditions and increasingly special theorems. Also, there should be many examples justifying all of this structure. The work I have been discussing in this talk is very general. There is still a good deal of general work left to be done; even the connections between infinite dimensional realizability and other areas is not yet fully developed (continuous time theory in particular). However, I suspect that this will go pretty quickly because a good deal of technique has been developed within mathematical physics and functional analysis which can be applied readily. Soon, and even now, we are confronted with the task of finding the physically interesting additional restrictions which constitute the next layers of specificity.

Let me indulge in a rambling discussion of some things that one might look for. Most of the systems I have much acquaintance with have frequency response functions which are meromorphic except possibly at points on the unit circle. These are called pseudo-meromorphic functions. I suspect that all transmission lines have such frequency response functions and quite possibly some engineer could tell me very quickly whether or not this is so. A mathematician's wave guide does not have such a scattering matrix, although the engineers approximation might.

It is clear from basic properties of pseudo-meromorphic functions and Chapter VIII, §2 [13], that the spectrum of A arising in the restricted shift realization for a matrix valued pseudo-meromorphic function $F(z)$ consists of isolated points located at the poles of $F(z)$ (reflected about the unit circle) union a closed subset of the unit circle. The interior points are all eigenvalues of A and the spectral space for A at these points consists entirely of generalized eigenvectors. More structure than just this which is gleaned from abstract considerations probably exists in practice. The examples which Lax-Phillips treat; c.f., Theorem 5.5 [11] (scattering off of obstacles) make one suspect that frequently say for a lossless transmission line, the spectrum of A will have exactly one accumulation point. Thus, the only spectrum of A will be point spectrum union one accumulation point. This is a much simpler situation that can arise in general (see Chapter VIII [13]) and one notes that even the weakest form of the Uniqueness Theorem when applied to state operators A and $\hat{A}$ having this property formally yields that the spectrum is the same. In addition, Chapter V, §3, Conjecture [10], suggests that certain critical operators may frequently be compact. I don't have much idea how much spectrum on A one could get on the unit circle when dealing with an RC line. That would be a nice thing to know.

The layer of structure just described may well be present in most situations where the frequency response function is pseudo-meromorphic. It has been exposed to view largely by work of Lax-Phillips and Nagy-Foias. I have no idea what the level of detail beyond this and alternative to this looks like. It seems to be territory which is uncharted by mathematicians. It seems like an area where, if communications were good, engineers could be of considerable assistance to mathematicians, in suggesting the right assumptions for producing an elegant and physically realistic theory.

REFERENCES

[1] A. V. Balakrisna, On the state space theory of linear systems, Jour. Math. Ann. and Appl. v. 14 (1966) 371-391.

[2] J. Baras and R. Brockett, (to appear).

[3] P. DeWilde, Roomy scattering matrix synthesis, Berkeley Technical Report, 1971.

[4] R. G. Douglas and J. W. Helton, Inner dilations of analytic matrix functions and Darlington synthesis, Acta Sci. Math. (Szeged) Tom 34 (1973).

[5] P. Fuhrmann, On realization of linear systems and applications to some questions of stability, (to appear).

[6] P. Fuhrmann, Realization theory in Hilbert space for a class of transfer functions (to appear).

[7] F. J. Helton and J. W. Helton, Scattering theory for computers and other non-linear systems (preprint available).

[8] J. W. Helton, Discrete time systems, operator models, and scattering theory, (to appear) Jour. Func. Analysis.

[9] J. W. Helton, The infinite dimensional optimal regulator and the algebraic Ricatti equation; a spectral factorization approach, (in preparation).

[10] P. D. Lax and R. S. Phillips, Scattering Theory, Academic Press, 1967.

[11] P. D. Lax and R. S. Phillips, Scattering theory for dissipative hyperbolic systems, Acta Sci. Math. (Szeged) Tom 34 (1973).

[12] M. S. Livsic, Operators oscillations and waves; open systems, AMS Monograph (1973).

[13] B-Sz. Nagy and C. Foias, Harmonic analysis of operators on Hilbert space, North Holland, Amsterdam, 1970.

[14] J. Rovnyak and M. Rosenblum, The factorization problem for non-negative operator functions, Bull. Amer. Math. Soc. 77 (1971) 287-318.

EXISTENCE OF OPTIMAL CONTROLS FOR NONLINEAR DISTRIBUTED PARAMETER SYSTEMS

MARSHALL SLEMROD
Department of Mathematics, Rensselaer Polytechnic Institute
Troy, New York 12181

INTRODUCTION

This research announcement attempts to provide a unified theory for the existence of optimal controls for dynamical control systems governed by nonlinear partial differential equations. We will consider problems of the time optimal and cost optimal nature. The main technique is to imbed the state equations into the theory of nonlinear evolution equations in Banach spaces as given in [1] and make use of certain approximation theorems presented there. Similar problems have been studied by J. L. Lions [2,3,4] and while our theory is in some sense more unified it does not at present include all the examples presented by Lions.

1. PRELIMINARY RESULTS

We present some of the basic results on existence of solutions to nonlinear partial differential equations in Banach space. Since we are interested in a unified theory for existence of optimal controls for controls systems governed by nonlinear partial differential equations it makes sense to make use of the available theory on existence of solutions to nonlinear evolution equations given in [1].

Let X be a real Banach space with norm $||\cdot||$. A subset A of X x X is in the class $\mathcal{A}(\omega)$ if for $\lambda > 0$ such that $\lambda\omega < 1$ and each pair $[x_i, y_i] \in A$, $i = 1,2$ we have

$$||(x_1+\lambda y_1) - (x_2+\lambda y_2)|| \geq (1-\lambda\omega)||x_1 - x_2||.$$

We define for subsets A and B of X x X and any real number α the following:

(i) $D(A) = \{x:[x,y] \in A \text{ for some } y\}$
(ii) $R(A) = \{y:[x,y] \in A \text{ for some } x\}$
(iii) $Ax = \{y:[x,y] \in A\}$ for $x \in D(A)$
(iv) $\alpha A = \{[x,\alpha y]:[x,y] \in A\}$
(v) $A^{-1} = \{[y,x]:[x,y] \in A\}$
(vi) $A + B = \{[x,y+z]:[x,y] \in A \text{ and } [x,z] \in B\}$
(vii) $AB = \{[x,z]: \text{for some } y\ [x,y] \in B \text{ and } [y,z] \in A\}$.

A is called *accretive* if $A \in \mathcal{A}(0)$. If λ is real, J_λ will denote the set $(J + \lambda A)^{-1}$ and $D_\lambda = D(J_\lambda) = R(J + \lambda A)$. It is not hard to see that for $A \in \mathcal{A}(\omega)$, $\lambda \geq 0$, $\lambda\omega < 1$, J_λ is actually a function (even though A viewed as an operator may be multivalued) and for $x,y \in D_\lambda$

$$||J_\lambda x - J_\lambda y|| \leq (1 - \lambda\omega)^{-1}||x - y||.$$

Now let T,ω denote real numbers, $T > 0$, and assume $A(t)$ satisfies

(A.1) $A(t) \in \mathcal{A}(\omega)$ for $0 \leq t \leq T$.

(A.2) $\overline{D(A(t))} = X$ is independent of t. (We choose, e.g., $D = D(A(0))$.

(A.3) $R(I + \lambda A(t)) = X$ for $0 \leq t \leq T$ and $0 < \lambda < \lambda_o$, where $\lambda_o > 0$, $\lambda_o\omega < 1$.

Let $J_\lambda(t) = (J + \lambda A(t))^{-1}$. The t-dependence of $A(t)$ will be restricted by the condition:

(C.1) There is a continuous function $f:[0,T] \to X$ and a monotone increasing function $L:[0,\infty] \to [0,\infty]$ such that

$$||J_\lambda(t)x - J_\lambda(\tau)x|| \leq ||f(t) - f(\tau)||L(||x||)$$

for $0 < \lambda < \lambda_o$, $0 \leq t,\tau \leq T$, and $x \in \overline{D}$.

It follows from conditions (A.1) to (A.3) that for each fixed τ, $0 \leq \tau \leq T$, there is a semigroup $S_\tau(t)$ on X associated with $A(\tau)$. See [5]. The condition (c.1) assumes only the continuity of f and corresponds roughly to the case in which $A(t)$ has the form $A(0) + B(t)$, where $B(t)x$ is well-behaved in x. In control theory this corresponds to the case when $B(t)$ represents the control in the coefficient of the zero order terms of the nonlinear differential operator $A(t)$. That is to say our control system will satisfy an equation of the form

$$\frac{du}{dt} + A(0)u + B(t)u \ni 0.$$

Also taking $B(t)x$ to be independent of x so that $-B(t)x=h(t)$ we can treat the quasi-autonomous case

$$\frac{du}{dt} + A(0)u \ni h(t).$$

This corresponds in applications to the case where $h(t)$ is a distributed control.

In the theory of nonlinear evolution equations the main interest is in showing the existence of evolution operators. An *evolution operator* $U(t,s)$ will be an operator $X \to X$ satisfying the relations

(i) $U(s,s) = I$ (the identity operator)

(ii) $U(t,s)U(s,r) = U(t,r)$ for $0 \le r \le s \le t \le T$.

(iii) $U(t,s)x$ is continuous in the pair (t,s) on the triangle $0 \le s \le t \le T$.

(iv) $||U(t,s)x - U(t,s)y|| \le e^{\omega(t-s)}||x - y||$ for $0 \le s \le t \le T$ and x,y in X.

One would hope that the evolution operator $U(t,s)x$ will provide in some sense a generalized solution to the Cauchy problem

$$\begin{aligned} \frac{du}{dt} + A(t)u &\ni 0 \\ u(s) &= x. \end{aligned} \tag{1.1}$$

This indeed is the case when $A(t)$ satisfies (A.1) to (A.3) and (C.1). This is brought out more clearly if we make the following definition. A function $u:[s,T] \to X$ is called a strong solution of (1.1) if:

(i) u is continuous on $[s,T]$ and $u(s) = x$,

(ii) u is absolutely continuous on compact subsets of (s,T),

(iii) u is differentiable a.e. on (s,T) and satisfies (1.1) a.e.

We can now state the following fundamental theorem.

Theorem:

[1] Let $A(t)$ satisfy (A.1), (A.2), (A.3), (C.1). Then there exists an evolution operator $U(t,s)$ on X. Furthermore if there exists a strong solution $u(t)$ to (1.1) then $u(t) = U(t,s)x$ for $0 \le s \le t \le T$.

Finally let us make one more definition. Let $A \varepsilon \mathcal{A}(\omega)$ and $R(I + \lambda A) = X$. For $x \varepsilon X$ set

$$|Ax| = \lim_{\lambda \downarrow 0} ||A_\lambda x||$$

where $A_\lambda = \lambda^{-1}(I - J_\lambda)$. Define

$$\hat{D}(A) = \{x \varepsilon X; |Ax| < \infty\} .$$

The set $\hat{D}(A)$ which has been studied in [6] is known as the generalized domain of the operator A.

We will use the concept of generalized domain to state the fundamental approximation theorem of [1]. It is not surprising that we will need a theorem of this type in proving the existence of optimal controls. The reason is simple. In the theory of optimal control one usually chooses a minimizing sequence of controls and then shows that this minimizing sequence possesses a subsequence which converges to the optimal control. Furthermore, one verifies that the sequence of states corresponding to the minimizing sequence of controls converges to the optimal state and that the optimal state and optimal control satisfy the original differential equation in some (possibly relaxed) sense.

Motivated by the above heuristic argument we will consider a family $A^{\beta}(t)$ of $\mathcal{A}(\omega)$ indexed by β, $0 \leq \beta < 1$, which satisfy the assumptions of Theorem 1. (In applications $A^{\beta}(t)$ may be considered in the form $A^{\beta}(t) = A(0) + B^{\beta}(t)$ which represents controls in the coefficients of the zero order terms of the nonlinear differential operator $A^{\beta}(t)$ or $A^{\beta}(t) = A(0)x - h^{\beta}(t)$ which represents distributed controls $h^{\beta}(t)$.) We let $U^{\beta}(t,s)$ be corresponding evolution operators. Suppose $A^{\beta}(t) \to A^{0}(t)$ (in some sense) as $\beta \to 0$. One expects that $U^{\beta}(t,s)$ will converge to $U^{0}(t,s)$. Our next result, Theorem 2, shows this is indeed the case under conditions which may be verified in many applications. We drop the index β if $\beta = 0$ below, e.g. $A^{0}(t) = A(t)$, $U^{0}(t,s) = U(t,s)$, etc.

Theorem 2

[1] Let $A^{\beta}(t)$ satisfy the assumptions of Theorem 1 uniformly in β, $0 \leq \beta < 1$ (i.e. we may take the same ω, λ_{o}, T, f, L for each $A^{\beta}(t)$). Let $J_{\lambda}^{\beta}(t) = (I + \lambda A^{\beta}(t))^{-1}$ and assume $\overline{D(A)^{\beta}(0))} = X$ and

$$\lim_{\beta \downarrow 0} J_{\lambda}^{\beta}(t)x = J_{\lambda}(t)x \quad \text{for } x \in \hat{D}(A(0)),$$

$0 < \lambda < \lambda_{o}$, $0 \leq t \leq T$. Let $U^{\beta}(t,s)$ be the evolution operators on X associated with $A^{\beta}(t)$ in the sense of Theorem 1. Then

$$\lim_{\beta \downarrow o} U^{\beta}(t,s)x = U(t,s)x$$

for each $x \in \overline{\hat{D}(A(0))}$, $0 \leq s \leq t \leq T$ and the limit is uniform in $t \in [s,T]$.

In applications to control theory we associate the sequence $\{A^{\beta}(t)\}$ as our minimizing subsequence of controls. The only apparent difficulty in applying Theorem 2 is to show that

$$\lim_{\beta \downarrow o} J_{\lambda}^{\beta}(t)x = J_{\lambda}(t)x \quad \text{for } x \in \hat{D}(A(0))$$

where A(t) will represent our "optimal control". A partial solution to this problem is given by the following result of

Brezis and Pazy [7] which is stated here in a simplified form.

Theorem 3

[7] Let $A \in \mathcal{A}(\omega)$ be single valued, closed, and satisfy $R(I + \lambda A) = \overline{X, D(A)} = X$. Let $A^\beta \in \mathcal{A}(\omega)$ be single valued and satisfy $R(I + \lambda A^\beta) = X$ for $0 < \lambda < \frac{1}{\omega}$. If $D(A^\beta) \supset D(A)$ then

$$\lim_{\beta \to 0} J_\lambda{}^\beta x = J_\lambda x \quad \text{for every } x \in X.$$

Combining Theorems 2 and 3 we immediately obtain

Corollary 1

Let $A^\beta(t)$, $0 \leq \beta < 1$, $A(t)$ be in $\mathcal{A}(\omega)$, single valued, $R(I+\lambda A^\beta(t)) = R(I + \lambda A(t)) = X$ for $0 < \lambda < \frac{1}{\omega}$, $0 \leq t \leq T$. Assume $A(t)$ is closed and $D(A^\beta(t)) = D(A(t)) = D(A(0))$, $\overline{D(A^\beta(0))} = \overline{D(A(0))} = X$ and $A^\beta(t)$ satisfy the assumptions of Theorem 1 uniformly in β. Let $U^\beta(t,s)$ be the evolution operator associated with $A^\beta(t)$, $U(t,s)$ be the evolution operator associated with $A(t)$ in the sense of Theorem 1. Then if $A^\beta(t)x \to A(t)x$ for every $x \in D(A)$ and $0 \leq t \leq T$ then

$$\lim_{\beta \downarrow 0} U^\beta(t,s)x = U(t,s)x$$

for each $x \in X$, $0 \leq s \leq t \leq T$, and the limit is uniform in $t \in [s,T]$.

Remark. The proof is a simple combination of Theorems 3 and 4 where we have used the fact that $\hat{D}(A(0)) \supset D(A(0))$ and $\overline{D(A(0))} = \overline{\gamma(A(0))} = X$. (See [1] or [6]).

2. TIME OPTIMAL CONTROL PROBLEM

Let X be a Banach space and K a closed set in X. Let $\mathcal{C}$ be a class of generators such that if $A(t) \in \mathcal{C}$ then

(i) $A(t)$ satisfies (A.1), (A.2), (A.3),

(ii) (C.1) is satisfied uniformly, i.e. same f, L, λ work for all $A(t) \in \mathcal{C}$,

(iii) $A(t)$ is single valued,

(iv) $D(A(t)) = D(A(0)) = D$(independent of A),

(v) $\overline{D} = X$.

The class $\mathcal{C}$ will be determined by the form of the state equation and the set of admissible controls.

Assume for some $\tau \in [0,T]$ and $A(t) \in \mathcal{C}$ then evolution operator $U(t,s)$ generated by $A(t)$ via Theorem 1 satisfies $U(\tau,0)x \in K$ for $x \in X$. Set τ_0 (the optimal time) = inf τ. We would like to show the existence of an "optimal generator" $A^*(t)$ in $\mathcal{C}$ such that $A^*(t)$ generates an "optimal evolution operator" $U^*(t,s)$ via Theorem 1 so that $U^*(\tau_o,0)x \in K$.

Let $\{\tau_\beta\}$ be a minimizing sequence of times so that $\tau_\beta \to \tau_o$ as $\beta \downarrow 0$. Let $\{A^\beta(t)\}$ be the associated sequence of generators.

Theorem 3

Under the above assumptions if there exists a subsequence of $\{A^\beta(t)\}$ (also denoted by $A^\beta(t)$) and $A^*(t) \in \mathcal{C}$ so that $A^\beta(t)x \to A^*(t)x$ for all $x \in D$ then the evolution operator $U^*(t,s)$ associated with $A^*(t)$ satisfies $U^*(\tau_o,0)x \in K$.

Proof By Corollary 1 $\lim_{\beta\downarrow 0} U^\beta(t,0)x = U^*(t,0)x$ uniformly in $t \in [0,T]$. For $x \in X$, $||U^\beta(\tau_\beta,0)x - U^*(\tau_o,\alpha)x|| \leq ||U^\beta(\tau_\beta,0)x - U^*(\tau_\beta,0)x|| + ||U^*(\tau_\beta,0)x - U^*(\tau_o,0)x||$. But since $\lim_{\beta\downarrow 0} U^\beta(t,0)x = U^*(t,0)x$ uniformly in $t \in [0,T]$ the first term in the right hand side of the above inequality goes to zero as $\beta \downarrow 0$. Since $U^*(t,0)x$ is continuous in t the second term in the inequality also goes to zero as $\beta \downarrow 0$ so that $U^*(\tau_o,0)x \in k$.

3. A COST OPTIMIZATION PROBLEM

Let X be a Banach space and K a closed set in X. Let $\mathcal{C}$ be a class of generators satisfying (i) - (v) of Section 2 which will be determined by the form of the state equation and set of admissible controls. Assume for some $\tau \in [0,T]$ and $A(t) \in \mathcal{C}$ the evolution operator $U(t,s)$ generated by $A(t)$ via Theorem 1 satisfies $U(\tau,0)x \in K$ for $x \in X$. Let F be a "cost functional" such that $F:C([0,\tau];X) \to R$ so that if $w_n \to w$ as $n \to \infty$ in $C([0,\tau];X)$ we have $F(w) \leq \liminf F(w_n)$. We can then pose the following problem.

Problem

Can we find an $A^*(t)$ in $\mathcal{C}$ generating an evolution operator $U^*(t,s)$ so that $F(U^*(\cdot,0)x) \leq F(U(\cdot,0)x)$ for all other evolution operators generated in the sense of Theorem 1 by $A(t)$ in $\mathcal{C}$ and where $U^*(\tau,0)x \in K$?

The problem may be resolved as follows. Let $\{A^\beta(t)\}$ be a minimizing sequence of generators in $\mathcal{C}$. Assume the corresponding evolution operators satisfy $U^\beta(\tau,0)x \in K$ and let $F_o = \inf_\beta F(U^\beta(\cdot,0)x)$. For convenience we order the minimizing sequence so that

$$\lim_{\beta\downarrow 0} F(U^{\beta}(\cdot,0)x) = F_o .$$

Theorem 4

Under the above assumptions if there exists a subsequence of $\{A^{\beta}(t)\}$ (also denoted by $\{A^{\beta}(t)\}$ and $A^*(t) \varepsilon \mathcal{C}$ so that $A^{\beta}(t)x \rightarrow A^{\alpha}(t)x$ for all $x \varepsilon D$ then the evolution operator $U^*(t,s)$ corresponding to $A^*(t)$ satisfies $U^*(\tau,0)x \varepsilon K$ and $F(U^*(\cdot,0)x) = F_o$.

Proof By Corollary 1 $\lim_{\beta\downarrow 0} U^{\beta}(t,0)x = U^*(t,0)x$ uniformly in $t \varepsilon [0,T]$. Now for $x \varepsilon X$ $||U^{\beta}(\tau,0)x - U^*(\tau,0)x|| \rightarrow 0$ as $\beta\downarrow 0$ so $U^*(\tau,0)x \varepsilon K$. Furthermore $F(U^*(\cdot,0)x) \leq \lim\inf F(U^{\beta}(\cdot,0)x) = F_o$ by the assumption on the cost functional F. Thus $F(U^*(\cdot,0)x) = F_o$ and the proof is complete.

4. AN EXAMPLE: NONLINEAR WAVE EQUATION WITH CONTROL IN THE DAMPING COEFFICIENT

Let Ω be a bounded domain in R^n with smooth boundary. Let $H_2(\Omega)$, $\overset{o}{H}_1(\Omega)$, $H_o(\Omega)$ denote the usually Sobolev spaces [2]. We will consider the nonlinear hyperbolic control system given by

$$\left.\begin{array}{r} \ddot{y} - \Delta y + \gamma(\dot{y}) + b(t)\dot{y} = 0, \ 0 \leq t \leq T, \\ y = 0 \quad \text{on } \partial\Omega \end{array}\right\} \qquad (4.1)$$

where Δ is the Laplacian, γ is continuous on $H_o(\Omega)$, $\gamma(0) = 0$ and γ satisfies the monotonicity condition

$$(\gamma(v) - \gamma(w), v - w)_{H_o(\Omega)} \geq 0.$$

The controls $b(t)$ lie in the admissible control set $V = \{$a set of nonnegative equilipschitzian functions on $[0,T] \rightarrow R$, i.e., $|b(t) - b(\tau)| \leq M|t - \tau|$ (M independent of b), $b(t) \geq 0\}$. By Ascoli-Arzela Thm. V is compact.

We write (4.1) as a system

$$\dot{y}_1 = y_2$$
$$\dot{y}_2 - \Delta y_1 + \gamma(y_2) + b(t)y_2 = 0$$

and $u = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$.

Let X (the state space) be given by $\overset{o}{H}_1(\Omega) \oplus H_o(\Omega)$ which endowed with "energy" inner product

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \quad w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$(v,w)_X = \int_{\Omega} \{\nabla v_1 \cdot \nabla w_1 + v_2 w_2\} dx.$$

Let us introduce the operators

$$A_1 = \begin{pmatrix} 0 & 1 \\ \Delta & 0 \end{pmatrix}, \qquad A_2(t) = \begin{pmatrix} 0 & 0 \\ 0 & \gamma + b(t) \end{pmatrix}$$

where $D(A_1) = H_2(\Omega) \cap \overset{\circ}{H}_1(\Omega) \oplus H_1^0(\Omega)$ and $D(A_2(t)) = X$. In this case it is well known that A_1 is the infinitesimal generator of a C_o contraction semigroup on X (the generalized solution to the wave equation with homogeneous boundary conditions). A_1 is closed on $D(A_1)$ and satisfies (A.1), (A.2), (A.3) where $\omega = 0$ (say by the Hille-Yosida-Phillips Theorem).

Define

$$A(t) = -A_1 + A_2(t)$$

$\mathcal{C}$ will be the set of generators $A(t)$ of this form where $b \in V$.

We now see our original control system (4.1) can be written as

$$\dot{u}(t) + A(t)u = 0 \tag{4.2}$$

where $D(A(t)) = D(A_1)$.

We must show $\mathcal{C}$ satisfies (i) - (vi) of section 2.

(i) For $A(t) \in \mathcal{C}$ integration by parts shows $A(t)$ is monotome in X and hence in $\mathcal{A}(0)$. $D(A(t)) = D(A_1)$. We will check $R(I + A(t)) = X$ later.

(ii) We will check (C.1) later.

(iii) $A(t)$ is certainly single valued.

(iv) $D(A(t)) = D(A_1)$ for all $A(t)$ in $\mathcal{C}$.

(v) $\overline{D(A_1)} = \overset{\circ}{H}_1(\Omega) \oplus H_o(\Omega) = X$

To check the range condition on $A(t) \in \mathcal{C}$ we will employ the following theorem of Webb [8].

Theorem

[8]. Let A be a closed, densely defined, linear accretive operator from a Banach space X to itself, $R(I + \lambda A) = X$ for some $\lambda > 0$. Let B be a continuous everywhere defined nonlinear accretive operator from X to itself. Then $A + B$ is accretive and $R(I + \lambda(A + B)) = x$ for some $\lambda > 0$.

Using the above theorem with $A = -A_1$ and $B = A_2(t)$ we see $R(I + \lambda A(t)) = X$ for some $\lambda > 0$. (Any $\lambda > 0$ is sufficient for our purposes since $\omega = 0$). Thus (i) is satisfied.

We must now check (ii). Let us define as usual $J_\lambda(t) = (I + \lambda A(t))^{-1}$. Then for $x \in X$, $\lambda > 0$ and $z(t) = J_\lambda(t)x$ we have

$$z(t) - z(\tau) = -\lambda(A(t)z(t) - A(\tau)z(\tau)).$$

Taking the inner product in X with $z(t) - z(\tau)$ we get

$$||z(t) - z(\tau)||^z_X = -\lambda(A_2(t)z(t) - A_2(\tau)z(\tau), z(t)-z(\tau))_X$$

since $A_1 = -A_1^*$ on X. From the monotonicity of γ on $H_o(\lambda)$ we get

$$||z(t)-z(\tau)||^2_X \leq -\lambda\left(\begin{pmatrix}0 & 0\\ 0 & b(t)\end{pmatrix} z(t) - \begin{pmatrix}0 & 0\\ 0 & b(t)\end{pmatrix} z(t) + \begin{pmatrix}0 & 0\\ 0 & b(t)\end{pmatrix} z(\tau) - \begin{pmatrix}0 & 0\\ 0 & b(\tau)\end{pmatrix} z(\tau),\ z(t) - z(\tau)\right)_X$$

Since $J_\lambda(t)0 = 0$ we $||z(t)||_X \leq ||x||_X$ and since $|b(t) - b(\tau)| \leq M|t - \tau|$ we have from the Schwarz inequality that

$$||J_\lambda(t)x - J_\lambda(\tau)x||_X \leq \lambda M|t - \tau|\ ||x||_X \text{ for all } x \in X$$

and M is independent of $b(t)$. So (ii) is satisfied.

Now let $\{b_\beta(t)\} \subset V$ be a minimizing sequence of controls for either the time optimal or cost optimal problems. Since V is compact $\{b_\beta(t)\}$ possesses a convergent subsequence also denoted by $\{b_\beta(t)\}$ so that $b_\beta(t) \to b^*(t) \in V$. ($b^*(t)$ is our optimal control). Denote

$$A^\beta(t) = -A_1 + \begin{pmatrix}0 & 0\\ 0 & \gamma+b_\beta(t)\end{pmatrix} \quad \text{and}$$

$$A^*(t) = -A_1 + \begin{pmatrix}0 & 0\\ 0 & \gamma+b^*(t)\end{pmatrix}.$$ We can now repeat the above arguments to show $A^*(t) \in \mathcal{C}$. It is obvious that $A^\beta(t)x \to A^*(t)x$ for $x \in D = D(A_1)$. Thus all the hypotheses of Theorems 3 and 4 are satisfied.

Conclusion

For control system (4.1) with admissible control set V the time optimal and cost optimal problems possess solutions.

REFERENCES

[1] M. G. Crandall and A. Pazy, Nonlinear evolution equations in Banach spaces, Israel J. Math. 11 (1972), 57-94.

[2] J. L. Lions, "Optimal control of systems governed by partial differential equations", Springer-Verlag, New York (1971).

[3] J. L. Lions, On some optimization problems for linear parabolic equations, in "Functional analysis and optimization" ed. E. R. Caianiello, Academic Press, New York (1966), 115-131.

[4] J. L. Lions, Optimization pur certaines classes d'equations

d'évolution non linéaire, Annali di Mat. LXXII (1966), 275-294.

[5] M. G. Crandall and T. M. Liggett, Generation of semigroups of nonlinear transformations on general Banach spaces, Amer. J. Math. 93, (1971), 265-298.

[6] M. G. Crandall, A generalized domain for semigroup generators, MRC Technical Summary Report #1189.

[7] H. Brezis and A. Pazy, Convergence and approximation of nonlinear operators in Banach spaces, J. Functional Analysis 9 (1972), 63-74.

[8] G. F. Webb, Continuous nonlinear perturbations of linear accretive operators in Banach spaces, J. Functional Analysis 10 (1972), 191-203.

A REALIZATION CONDITION FOR INFINITE DIMENSIONAL BILINEAR SYSTEMS*

GIORGIO KOCH
Istituto di Automatica
University of Rome
Rome, Italy

ABSTRACT

The paper deals with infinite dimensional bilinear systems. Apparently it is the first contribution in this area, except for an earlier paper [4] on optimal control problems.

First, the state space representation of an infinite dimensional bilinear system is given. Then the problem is dealt with of the connection between the latter and the input-output map. The Volterra series expansion of the zero-state response is derived and conversely a necessary and sufficient condition is obtained for an input-output relation to be realizable by means of a bilinear system.

1. INTRODUCTION

In the last few years, the work of Mohler and his co-authors (see for instance [1]) stimulated researchers in the field of bilinear systems, i.e., systems with products between state and input variables.

The theoretical and applicative interest of these systems originates from their character of simplest variable structure systems. Therefore, they are close enough to linear systems to present a number of properties which may be analyzed through a generalization from the linear case, and at the same time they may account for a control of the input signal on the structure of the system, thus suitably modeling several relevant phenomena in physics, biology, socio-economics and ecology.

For a comprehensive exposition of the theory and applications of bilinear systems, see [2]. Several papers on bilinear systems are also contained in [3]. As it appears from [2], and as far as the author is aware, practically all research works on bilinear systems consider the finite dimensional case; there is a single paper [4] on infinite dimensional bilinear systems which is concerned with an optimal control problem.

The present paper is intended to achieve some preliminary results in the theory of infinite dimensional bilinear systems. Attention is focused on the representation of these systems by their state equations, or by input-output relations. First, the state space representation is given. Then, paralleling previous results obtained in the finite dimensional case [5, 6], we investigate the connection between the state space representation and the input-output one. Given the former one, a series expansion of the output in terms of the initial state and the input is derived, which in particular yields a Volterra-type series expansion of the zero-state response.

Conversely, a necessary and sufficient condition is obtained for a zero-state input-output relation to be realizable by means of an infinite dimensional bilinear system. This condition appears as an extension to the

---

*This work was supported in part by Italian National Research Council (CNR)

infinite dimensional case of a similar condition proved in [6, 7]. Essentially, the condition is the factorization of the kernels of the input-output map. To prove it, advantage is taken from recent results on the realization problem for infinite dimensional linear systems [8, 9].

The theory of infinite dimensional bilinear systems, although at its very beginning, appears quite promising and deserves to be the object of future research.

2. INFINITE DIMENSIONAL BILINEAR SYSTEMS

In this paper we consider time invariant infinite dimensional bilinear systems defined by the following differential equation

$$\dot{x}(t) = Ax(t) + N(x(t), u(t)) + Bu(t), \quad x(0) = \hat{x} \tag{2.1}$$

$$y(t) = Cx(t) \tag{2.2}$$

As the state space H we take the Hilbert space $L_2[0,\infty)$. The input u is in R^p for each $t \geq 0$; moreover u is assumed to be measurable with respect to t, and uniformly bounded on any finite time interval. The output y is in R^q for each $t \geq 0$

A is the infinitesimal generator of a strongly continuous semigroup of linear bounded operators S(t) mapping H into H. B and C are linear bounded operators mapping respectively R^p into H, and H into R^q. Finally, N is a bilinear operator mapping H x R^p into H: for each u(t) in R^p, $N(\cdot,u(t))$ is a linear bounded operator from H into H, and for each x(t) in H, $N(x(t),\cdot)$ is a linear bounded operator from R^p into H.

Also, there exist p linear bounded operators N_i, $i = 1, 2, \ldots, p$ from H into H such that if we denote by $u_i(t)$, $i = 1, 2, \ldots, p$ the i-th component of u(t) we can also define N as follows

$$N(x(t), u(t)) = \sum_{1}^{p} N_i x(t)\, u_i(t) \tag{2.3}$$

or in a more compact form

$$N(x(t), u(t)) = \bar{N}[x(t) \otimes u(t)] \tag{2.4}$$

where

$$\bar{N} = [N_1\ N_2\ \ldots\ N_p] \tag{2.5}$$

and the tensor product $x(t) \otimes u(t)$ is defined in the usual way

$$x(t) \otimes u(t) = \begin{bmatrix} x(t)\, u_1(t) \\ x(t)\, u_2(t) \\ \vdots \\ x(t)\, u_p(t) \end{bmatrix} \tag{2.6}$$

Note that (2.1) may be rewritten as

$$\dot{x}(t) = [A + N(\cdot, u(t))]\, x(t) + Bu(t) \tag{2.7}$$

or

$$\dot{x}(t) = Ax(t) + [N(x(t),\cdot) + B]\, u(t) \tag{2.8}$$

so that the variable structure character is enlightened.

The following hypotheses will be made:

H1) B and $\overline{N}$ have range in the domain of A, $\mathscr{D}(A)$.

H2) The operators $S(t)B$ and $S(t)\overline{N}$ are such that

$$||\dot{S}(t)B|| \leq \sigma_1 e^{-\sigma_2 t}, \quad \text{for } t \to \infty \tag{2.9}$$

$$||\dot{S}(t)\overline{N}|| \leq \sigma_3 e^{-\sigma_4 t}, \quad \text{for } t \to \infty \tag{2.10}$$

where σ_1, σ_2, σ_3, σ_4 are suitable positive constants.

H1) guarantees that (2.1) has a unique differentiable solution for each $\hat{x}$ in $\mathscr{D}(A)$. If H1) does not hold, then a solution of (2.1) exists only in a weaker sense [8]. Note that $\mathscr{D}(A)$ is dense in H so that any element in H is the limit of a sequence of elements in $\mathscr{D}(A)$.

H2) is a sort of stability assumption, needed in the proof of Theorem 2, below. In this paper, systems for which H2) hold will be called stable.

Theorem 1

Let y be the output of a time invariant bilinear system (2.1), (2.2), with given initial state $\hat{x}$ and input u. Let S(t) be the strongly continuous semigroup of operators generated by A. Then on any finite time interval [0,T], y has the following series expansion

$$y(t) = CS(t)\hat{x} + \sum_{i=1}^{\infty} \int_0^t \cdots \int_0^{t_{i-1}} CS(t-t_1)N(S(t_1-t_2)(N(S(t_2-t_3)\ldots$$
$$S(t_{i-1}-t_i)[N(S(t_i)\hat{x},\cdot) + B]u(t_i),u(t_{i-1})),\ldots u(t_1))dt_i\ldots dt_1 \tag{2.11}$$

and the series converges uniformly in [0,T].

Proof. Let us define in H the sequence $\{x_i\}$ of solutions of the equation

$$\dot{x}_o(t) = Ax_o(t) + Bu(t) \tag{2.12}$$

$$\dot{x}_i(t) = Ax_i(t) + N(x_{i-1}(t),\, u(t)) + Bu(t), \quad i = 1, 2, \ldots \tag{2.13}$$

with initial conditions $x_i(0) = \hat{x}$, $i = 0, 1, 2, \ldots$. The (unique) solutions of the above equations have the form

$$x_o(t) = S(t)\hat{x} + \int_0^t S(t-t_1)Bu(t_1)dt_1 \tag{2.14}$$

$$x_i(t) = S(t)\hat{x} + \int_0^t S(t-t_1)[N(x_{i-1}(t_1),u(t_1))+Bu(t_1)]dt_1, \quad i = 1, 2, \ldots \tag{2.15}$$

while by integrating (2.1) we get for its solution x

$$x(t) = S(t)\hat{x} + \int_o^t S(t-t_1)[N(x(t_1),u(t_1)) + Bu(t_1)]dt_1 \qquad (2.16)$$

Hence

$$x(t)-x_i(t) = \int_o^t S(t-t_1)N(x(t_1)-x_{i-1}(t_1),u(t_1))dt_1, \quad i=1,2,\ldots \qquad (2.17)$$

Because of the assumptions on S(t), N, u, we have

$$||S(t)|| \leq \zeta_1 e^{t\zeta_2}, \quad \forall\, t \,\varepsilon\, [0,T] \qquad (2.18)$$

$$||N(x(t),u(t))|| \leq \nu\, ||x||, \quad \forall\, t \,\varepsilon\, [0,T]$$

where ζ_1, ζ_2, ν are suitable constants. Therefore a positive constant M exists, such that

$$||x(t)-x_i(t)|| \leq \int_o^t M||x(t_1)-x_{i-1}(t_1)||dt_1, \; i = 1,\, 2,\, \ldots,\, t \,\varepsilon\, [0,T] \qquad (2.19)$$

By iteratively substituting (2.17) in (2.19), we obtain

$$||x(t)-x_i(t)|| \leq M^i\int_o^t\int_o^{t_1}\cdots\int_o^{t_{i-1}} ||x(t_i)-x_o(t_i)||dt_i dt_{i-1}\cdots dt_1 \qquad (2.20)$$

Now $x(t)-x_o(t)$ is uniformly bounded (continuous) in [0,T]; therefore a constant k exists such that

$$||x(t)-x_o(t)|| \leq k, \quad \forall\, t \,\varepsilon\, [0,T] \qquad (2.21)$$

and (2.20) yields

$$||x(t)-x_i(t)|| \leq k\,\frac{M^iT^i}{i!}, \quad \forall\, t \,\varepsilon\, [0,T], \quad i = 1,\, 2,\, \ldots \qquad (2.22)$$

Consequently the sequence $\{x_i(t)\}$ converges to x(t) uniformly with respect to t in the sense of the norm of H. It follows that x(t) may be written as

$$x(t) = x_o(t) + \sum_{i=1}^{\infty} [x_i(t)-x_{i-1}(t)] \qquad (2.23)$$

where x_o is given by (2.14) and (defining: $x_{-1} \overset{\Delta}{=} 0$)

$$x_i(t)-x_{i-1}(t) = \int_o^t S(t-t_1)N(x_{i-1}(t_1)-x_{i-2}(t_1),u(t_1))dt_1, \; i=1,2,\ldots \qquad (2.24)$$

Recalling now (2.2), we easily get the expansion (2.11).

As it was natural to expect, expansion (2.11) appears to be quite similar to (and enjoys the same properties of) the analogous expansion [2, 5] in the finite dimensional case, with the semigroup S(t) taking the place of the transition matrix.

Let us now focus the attention on the zero-state response to which (2.11) reduces for $\hat{x} = 0$. To write it in a more compact form, we define a special matrix product.

Let R be a $m \times nr$ matrix, and S an $n \times s$ matrix. Then we define

$$R\square S = [R_1 \; R_2 \; \ldots R_r]\square S \triangleq [R_1 S \; R_2 S \; \ldots \; R_r S] \tag{2.25}$$

The same notation will also be used for matrices of operators $[R_1 \; R_2 \ldots \; R_r]$ acting on an element S in their domain.

Having defined the products $\square$ and $\otimes$, similarly to the finite dimensional case [5], the zero-state response can be given the form

$$y(t) = \sum_{i=1}^{\infty} \underbrace{\int_0^t \cdots \int_0^t}_{i\ \text{times}} W_i(t_1,t_2,\ldots,t_i)[u(t-t_1)\otimes u(t-t_2)\otimes\ldots\otimes u(t-t_i)]dt_1\ldots dt_i \tag{2.26}$$

where the $q \times p^i$ symmetrical kernels W_i are defined in the following way

$$W_1(t) = CS(t)B \tag{2.27}$$

$$W_i(t_1,t_2,\ldots,t_i) = \frac{1}{i!}\sum{}^{*} V_i(t_1,t_2,\ldots,t_i), \; i = 2, 3, \ldots \tag{2.28}$$

$$V_i(t_1,t_2,\ldots,t_i) = [CS(t_i)\bar{N}]\square[S(t_{i-1}-t_i)\bar{N}]\square\ldots\square[S(t_1-t_2)B] \quad \text{for } t_1 \geq t_2 \geq \ldots \geq t_i \tag{2.29}$$

$$= 0 \quad \text{otherwise}$$

In (2.28) $\sum^{*}$ means summation over all the $i!$ permutations of $t_1,t_2,\ldots,t_i$.

3. THE REALIZATION PROBLEM

In the previous section we deduced the input-output relation from the state equation of a bilinear system, and in particular we got the Volterra series expansion for the zero-state response.

The converse problem (realization problem) is also of noticeable interest: whether it is possible to get a prescribed state space representation (i.e., of the type (2.1),(2.2)) back starting from the knowledge of the (kernels of the) zero-state response.

The realization problem for finite dimensional bilinear systems has already been extensively investigated in [6, 7]. Thus we have a number of results which possibly could be conveniently extended to the infinite dimensional case.

In particular, a necessary and sufficient condition is known [7] for an infinite sequence of symmetrical kernels to be realizable by a finite dimensional bilinear system. The condition essentially is the factorizability of the kernels by means of three matrix functions with proper rational Laplace transform.

By relaxing this assumption we get the equivalent condition in the infinite dimensional case: the proof of that may be considered as the main result of this paper and a first contribution toward a complete realization theory for infinite dimensional bilinear systems.

Theorem 2

A necessary and sufficient condition for a system defined by a zero-state input-output relation of the type (2.26), in which W_i is a sequence of $q \times p^i$ symmetrical kernels, to admit a state space representation by means of a time invariant stable bilinear system of the type (2.1), (2.2), is that:

a) the first kernel W_1 satisfies the following conditions:

a1) W_1 is differentiable with respect to t;

a2) $||\dot{W}_1(t)|| \leq \mu_1 e^{-t\mu_2}$, for $t \to \infty$ where μ_1, μ_2 are suitable positive constants;

b) there exist three sets of linear bounded operators $\{F(t)\}$, $\{G(t)\}$, $\{L(t)\}$, $t \geq 0$, mapping respectively H^p into H, R^p into H, H^p into R^q, and satisfying the following conditions:

b1) they are strongly differentiable with respect to t on their domain;

b2) $\max (||\dot{F}(t)||, ||\dot{G}(t)||, ||\dot{L}(t)||) \leq \mu_3 e^{-t\mu_4}$, for $t \to \infty$ where μ_3, μ_4 are suitable positive constants;*

b3) $W_i(t_1,t_2,\ldots,t_i) = L(t_i)\square F(t_{i-1}-t_i)\square\ldots F(t_2-t_3)\square G(t_1-t_2)$, $i=2,3,\ldots$ on the set $\{t_1 \geq t_2 \geq \ldots \geq t_i\}$

Proof. Sufficiency. Given $W_1(t)$ and the three sets of operators $\{F(t)\}$, $\{G(t)\}$, $\{L(t)\}$, let us form the matrix of operators:

$$P(t) = \begin{bmatrix} W_1(t) & L(t) \\ G(t) & F(t) \end{bmatrix} \tag{3.1}$$

which obviously maps $R^p \times H^p$ into $R^q \times H$. Because of assumptions a1)-b2), it appears that $P(t)$ is linear, bounded and strongly differentiable on $R^p \times H^p$, with

$$||\dot{P}(t)|| \leq \pi_1 e^{-t\pi_2}, \quad \text{for } t \to \infty \tag{3.2}$$

where π_1, π_2 are suitable positive constants. $P(t)$ may be looked at as the weighting operator of a stable linear system, with input $\tilde{u}(t)$ in $R^p \times H^p$ and output $\tilde{y}(t)$ in $R^q \times H$; its zero-state response mapping is defined by

$$\tilde{y}(t) = \int_0^t P(t-s)\tilde{u}(s)ds \tag{3.3}$$

A consequence of (3.2) is that for each z in $R^p \times H^p$, and for each $t \geq 0$, $\dot{P}(t)z$ is an element in $R^q \times H$; as t ranges on $[0,\infty]$, $\dot{P}(\cdot)z$ is an element in the Hilbert space $\tilde{H}$ of functions of t with range in $R^q \times H$, equipped with the inner product

$$(f,g) \overset{\Delta}{=} \int_0^\infty [f(t), g(t)]dt, \quad f, g \in \tilde{H} \tag{3.4}$$

where [,] denotes the usual inner product in $R^q \times H$.

---

*If $\{F(t)\}$ is strongly differentiable, by a consequence of the Uniform Boundedness Principle, for each $t \geq 0$ the derivative $\dot{F}(t)$ may be defined as a linear bounded function; similarly for $\{G(t)\}$ and $\{L(t)\}$.

Now, given a positive real number h, define

$$\tilde{B}(\tau) = -h \sum_{0}^{\infty}{}_{i}\, \dot{P}(\tau + ih), \quad \tau \geq 0 \tag{3.5}$$

Due to (3.2) $\tilde{B}(\tau)$ is well defined since the series in (3.5) converges strongly for each $\tau > 0$; moreover, if B denotes the function $\tilde{B}(\cdot)$, we see that for each z in $R^p \times H^p$, $\tilde{B}z$ is an element of $\tilde{H}$ (actually, as a consequence of the Uniform Boundedness Principle, $\tilde{B}$ is a linear bounded function which maps $R^p \times H^p$ into $\tilde{H}$).

Note that (3.5) may be inverted as follows

$$P(t) = \frac{1}{h}\int_{0}^{h} \tilde{B}(t+\tau)d\tau, \quad t \geq 0 \tag{3.6}$$

Now let $\tilde{C}$ denote a linear bounded operator from $\tilde{H}$ into $R^q \times H$, defined by

$$\tilde{C}\tilde{x} = \frac{1}{h}\int_{0}^{h} \tilde{x}(\tau)d\tau, \quad \tilde{x} \in \tilde{H} \tag{3.7}$$

Finally, let $\tilde{A}$ denote the infinitesimal generator of the strongly continuous semigroup $\tilde{S}(t)$ of shift operators on $\tilde{H}$, defined as follows

$$\tilde{S}(t)\tilde{x} = \xi, \quad \xi(\tau) = \tilde{x}(t+\tau), \quad \tilde{x} \in \tilde{H}, \quad t \geq 0, \quad \tau \in [0,\infty) \tag{3.8}$$

Now it is possible to show that the zero-state input-output relation (3.3) admits the following state space representation [9]

$$\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}\tilde{u}(t), \quad \tilde{x}(0) = 0 \tag{3.9}$$

$$\tilde{y}(t) = \tilde{C}\tilde{x}(t) \tag{3.10}$$

where we take $\tilde{H}$ as state space. Indeed, from (3.9) we may solve for $\tilde{x}(t)$

$$\tilde{x}(t) = \int_{0}^{t} \tilde{S}(t-\sigma)\tilde{B}\tilde{u}(\sigma)d\sigma \tag{3.11}$$

and

$$\begin{aligned}\tilde{y}(t) = \tilde{C}\tilde{x}(t) &= \int_{0}^{t} \frac{1}{h}\int_{0}^{h} \tilde{S}(t-\sigma)\tilde{B}(\tau)u(\sigma)d\sigma d\tau \\ &= \int_{0}^{t} P(t-\sigma)\tilde{u}(\sigma)d\sigma \end{aligned} \tag{3.12}$$

Consequently, we get the equation

$$P(t) = \tilde{C}\tilde{S}(t)\tilde{B} \tag{3.13}$$

Now let $\tilde{C}$ and $\tilde{B}$ be partitioned as follows

$$\tilde{C} = \begin{bmatrix} \tilde{C}_o \\ \tilde{C}_1 \end{bmatrix}, \quad \tilde{B} = [\tilde{B}_o \;\; \tilde{B}_1 \; \ldots \; \tilde{B}_p] \tag{3.14}$$

where $\tilde{C}_o$ and $\tilde{C}_1$ have range respectively in R^q and H, while $\tilde{B}_o$ and $\tilde{B}_i$, $i=1,2,\ldots,p$ have domain respectively in R^p and H. Comparing (3.13), (3.14), with (3.1) leads to the following correspondences

$$W_1(t) = \tilde{C}_o\tilde{S}(t)\tilde{B}_o \tag{3.15}$$

$$L(t) = \tilde{C}_o\tilde{S}(t)[\tilde{B}_1 \quad \tilde{B}_2 \ \dots \ \tilde{B}_p] \tag{3.16}$$

$$G(t) = \tilde{C}_1\tilde{S}(t)\tilde{B}_o \tag{3.17}$$

$$F(t) = \tilde{C}_1\tilde{S}(t)[\tilde{B}_1 \quad \tilde{B}_2 \ \dots \ \tilde{B}_p] \tag{3.18}$$

Note that from (3.6) it follows that $\tilde{B}$ is strongly differentiable, so that its range is in $\mathscr{D}(\tilde{A})$. The same obviously holds for the ranges of $\tilde{B}_i$, $i = 0, 1, \dots, p$.

Now let T be the isometric isomorphism mapping $\tilde{H}$ onto H, and consider the following bilinear system with state space H

$$\dot{x}(t) = A_o x(t) + N_o(x(t), u(t)) + B_o u(t), \quad x(0) = 0 \tag{3.19}$$

$$y(t) = C_o x(t) \tag{3.20}$$

where u(t) is in R^p, y(t) is in R^q, $A_o = T\tilde{A}T^{-1}$ is the infinitesimal generator of the shift semigroup on H, $B_o = T\tilde{B}_o$, $C_o = \tilde{C}_o T^{-1}$, and N_o is defined as in (2.4) through the linear bounded operator $\overline{N}_o$ from H^p into H

$$\overline{N}_o = [T\tilde{B}_1 \quad T\tilde{B}_2 \ \dots \ T\tilde{B}_p]\square C_1 T^{-1} \tag{3.21}$$

It now appears that H1) is satisfied for the system (3.19), (3.20). If we recall now the assumptions a1)-b2), we easily verify that the above defined bilinear system is stable.

Moreover, by assumption b3) the equations (2.27), (2.28), (2.29) for the given kernels W_i, i=1,2,... are satisfied on the set $\{t_1 \geq t_2 \geq \dots \geq t_i\}$; and being the kernels symmetric by assumption, the same equations are satisfied for any value of the variables $t_1, t_2, \dots, t_i$.

Necessity. This is almost trivial. Let (2.1), (2.2) define a time invariant stable bilinear system. Then its symmetric kernels W_i are given by (2.27)-(2.29), and a1) is satisfied. Choose

$$L(t) = CS(t)\overline{N} \tag{3.22}$$

$$G(t) = S(t)B \tag{3.23}$$

$$F(t) = S(t)\overline{N} \tag{3.24}$$

Since B and $\overline{N}$ have range in the domain of A, $\dot{W}_1(t)$, $\dot{L}(t)$, $\dot{G}(t)$, $\dot{F}(t)$ may be properly defined, and assumption H2) implies a2), b2). Finally, b3) is trivially satisfied.

Remark. For a theory of factorizable sequences, see [6].

Remark. Theorem 2 is still valid even if in a2), b2), μ_2 and μ_4 assume negative values, provided we drop the requirement of stability for the bilinear system. Obviously, the proof should be conveniently modified, along the line followed in [8].

Remark. In the case in which we constrain the range of $\overline{N}$ to be of finite dimensions, (say m_o), the same necessary and sufficient condition holds, with F(t), G(t), L(t), absolutely continuous matrices respectively of dimensions $m \times mp$, $m \times p$, $q \times mp$, with $m \geq m_0$.

Finally, if the bilinear system has a finite dimensional state space, it may be proved [6, 7] that a necessary and sufficient condition for realizability is the same as above, plus the condition of proper rational Laplace transform for $W_1(t)$, F(t), G(t), L(t).

ACKNOWLEDGEMENT

The author thanks Professor A. Ruberti for having encouraged him in pursuing this research.

REFERENCES

1. R. R. Mohler, "A Common Model in Socio-Economics, Ecology and Physiology," Proc. Int. Conf. on Systems, Man and Cybernetics, Anaheim, California, October 1971.
2. C. Bruni, G. DiPillo, G. Koch, "Bilinear Systems: An Appealing Class of 'Nearly Linear' Systems in Theory and Applications," IEEE Trans. on Automatic Control, to appear.
3. R. R. Mohler, A. Ruberti (eds.), Theory and Applications of Variable Structure Systems, Academic Press, 1972.
4. D. L. Briggs, C. N. Shen, "Switching Analysis for Constrained Bilinear Distributed Parameter System with Applications," Trans. ASME, J. of Basic Engineering, Vol. 91, 2, June 1969.
5. C. Bruni, G. DiPillo, G. Koch, "On the Mathematical Models of Bilinear Systems," Ricerche di Automatica, Vol. 2, 1, 1971.
6. P. d'Alessandro, A. Isidori, A. Ruberti, "Lectures on Bilinear Systems Theory," Notes for a course held at C.I.S.M., Udine, Italy, Springer Verlag, 1972.
7. P. d'Alessandro, A. Isidori, A. Ruberti, "Realization and Structure Theory of Bilinear Dynamical Systems," Reports 204 and 216, Istituto di Automatica, Univ. of Rome, Italy, 1972. SIAM J. Control, to appear.
8. A. V. Balakrishnan, Introduction to Optimization Theory in a Hilbert Space, Springer Verlag, 1972.
9. A. V. Balakrishnan, "System Theory and Stochastic Optimization," NATO Advanced Institute on Network and Signal Theory, Bournemouth, England, September 1972.

POPULATION PROJECTIONS AND ECONOMIC INTERACTIONS

WILLIAM D. O'NEILL
University of Illinois at Chicago Circle
Chicago, Illinois 60680

ABSTRACT

Population projection via a stochastic parameter model is examined. The model structure is developed in the context of data for the U.S. population. Under stochastic assumptions on the norm of the population projection matrix, a strong convergence and a weak convergence theorem are proved with the result that the stochastic assumptions are sufficient conditions for asymptotic, stochastic stability of the system null solution. A possible mechanism for controlling the expected values of the projection matrix is proposed with the intent of stabilizing population via fertility sensitive economic variables.

CONTEXT OF THE BASIC MODEL

All subsequent results of this paper depend on the structure and assumptions inherent in a basic recurrence model of demography. This section will develop the basic model in sufficient detail for later sections but for more detailed aspects of data smoothing, approximations and different forms of the model; for example as an integral, the reader is referred to the classic work of Keyfitz[1].

Two initial assumptions to be made are that the population to be described is closed to migration and the parameters of the model do not change over time. The migration assumption is never exactly true in practice but the errors made in neglecting migration seem tolerable relative to the mathematical inconveniences incurred by accounting for it. The stationary assumption will be removed after the model structure emerges. The model is female specific in every detail. All dependent variables and parameters refer exclusively to female numbers unless explicit note is made to the contrary. All data cited is for the U.S. and taken from Keyfitz and Flieger[2]. The argument t and index i will refer to, respectively, time and age and will range over the positive integers with each increment in t and i representing 5 years real time and 5 years of age.

MODEL STRUCTURE-EFFECTS OF MORTALITY

A general census measures, among other things, the numbers of living people in various age groups.

Definition 1.

$K_i(t)$; $i=0,1,2,\dots,17$ is the number of people alive at time $t\geq 0$ in age groups $0,1,2,\dots,17$.

This definition categorizes those alive after $t=0$ between ages 0 and 85+ years into groups of span $5i$ to $5i+4$ years. Thus all living people between ages 25 and 29 years inclusive are grouped under $i=5$.

Definition 2.

$m_{i,i-1}$; $i=1,2,\dots,17$ is the relative frequency with which people of age group $i-1$ live to age group i.

From definitions 1 and 2

$$K_i(t+1) = m_{i,i-1}K_{i-1}(t) \; ; i=1,2,\dots,17. \tag{1}$$

Equation (1) uses the stationary assumption in that it assumes the causes of mortality do not change over one increment of time. The calculation of of $m_{i,i-1}$ from census data is via the life table constructed on the basis of reported deaths and population for the given time t. Details can be found in [1]. It should be clear however that with or without stationarity $0 \leq m_{i,i-1} \leq 1$ for all i.

Equation (1) can project the population already alive at t=0 given the starting population by age groups with the exception of those for which i=0, i.e. newborns.

MODEL STRUCTURE-EFFECTS OF NATALITY

Definition 3.

$m_{o,i}$; i=1,2,...,17 is the expected number of births in the period t to t+1 to a person in age class i.

This definition completes the structure of the basic model since

$$K_o(t+1) = \sum_{j=0}^{17} m_{o,j} K_j(t) \qquad (2)$$

combined with (1) results in the matrix model

$$K(t+1) = LK(t) \qquad (3)$$

where L is a stationary matrix (the population projection matrix) having a first row composed of the natality parameters, $m_{o,i}$; i=0,1,2,..,17, and a subdiagonal of mortality parameters, $m_{i,i-1}$; i=1,2,...,17 and zeroes elsewhere.

MODEL STRUCTURE-GENERAL COMMENTS AND DATA CONSIDERATIONS

The structure of the L matrix is a logical consequence of definitions 1 through 3 but it is obvious the stationarity of L leads to simplistic population projections. However (3) can be made a reasonable short term projection model, e.g. over two future time periods, if the L matrix is updated at census periods which occur every two time periods. In this sense only the structure of the L matrix remains fixed in population projections and new data is considered with the old to improve the projections.

For stability analysis one need only consider the natality parameter submatrix of L. This submatrix results from the observation that $m_{o,i}=0$ for i=0 and i≥11 since people of age 0 to 4 and those greater than age 55 can be expected from biological considerations to bear few children. Thus

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline B & C \end{array}\right] \qquad (4)$$

where the dimension of the square, maternity matrix M is determined by the position of the first right-most 0 in the first row of L. Because of its importance in demography the elements of M have been calculated and tabulated for the U.S. population for every 5 years since 1920 using census data and also by calculating M retrospectively for non-census years using census data on age distributions [2]. Representative elements of M are shown in Table 1. Figure 1 is typical of how the elements of M have varied historically. The figure suggests that the subdiagonal of M has stabilized over the past two decades at constant values close to unity while the first row elements vary somewhat randomly over time. Such behavior is not unexpected considering that the subdiagonal elements probably are influenced to a great extent by the average quality and level of medical services available to the U.S. population while the first row elements are subject to conscious, many faceted, decision processes regarding family size made by people in the childbearing ages.

THE NONSTATIONARY MATERNITY SUBMATRIX

In this section the maternity submatrix of L will be assumed to be time varying. From M(t) will be derived an equivalent difference equation with time varying coefficients and from this difference equation an alternative time varying matrix representation that will be used in the stochastic analysis. The dimension of the maternity submatrix extends only over the reproductive period, i.e. for $0 \leq i \leq 10$. Therefore we define the maternity vector k(t) as

$$k(t) = \mathrm{col}[K_o(t), K_1(t), \ldots, K_{10}(t)].$$

Using the matrix partition in (4) we have

$$k(t+1) = M(t)k(t) \tag{5}$$

Equation (5) is equivalent to the linear time varying difference equation

$$K_o(t) = \sum_{j=1}^{10} m_{o,j}(t-1)K_o(t-j) \prod_{i=1}^{j} m_{i-1}(t+i-j-1) \tag{6}$$

Definition 4.

The time functions g_j given by:

$$g_j(t-1,\ldots,t-j+1) = m_{o,j}(t-1) \prod_{i=1}^{j} m_{i,i-1}(t+i-j-1) \;;\; j=1,2,\ldots,10$$

and considered as functions of j are called the net maternity function of the population.

With definition 4, equation (6) and the definition

$$x_j(t) = K_o(t-j) \tag{7}$$

for the jth element of the vector x the final form of the maternity submodel is

$$x(t+1) = A(t,t-1,\ldots,t-10)x(t). \tag{8}$$

The A matrix has a first row composed of elements of the net maternity function, a subdiagonal of ones and all other elements are zero. We note however that the lagged time dependence of A arises solely from the time dependence assumed in the mortality parameters of the original L matrix. This is evident from equation (6). As shown in the previous section data for the U.S. would indicate these parameters have stabilized near unity. Thus in the sequel we can assume that A depends on t alone.

Definition 5.

If the net maternity function depends on t alone then n(t), given by

$$n(t) = \sum_{j=1}^{10} g_j(t)$$

is called the net reproduction rate at time t.

This definition is particularly useful for the system under consideration when the proper norm is used. If B is any matrix of elements b_{ij} we will take

$$|B| = \max_i \sum_j |b_{ij}| \;. \tag{9}$$

From this norm the system matrix of equation (8) with A a function of t alone satisfies

$$|A(t)| = 1 + [n(t)-1][\mathrm{sgn}(n(t)-1)]. \tag{10}$$

THE STOCHASTIC MATERNITY MODEL

In this section certain assumptions are made concerning net reproduction rate and asymptotic results derived. Since it is an historical fact that n(t) has been notoriously unpredictable it would be natural to assume it to be a random variable and determine its probability density function from data. However, the existence of less than 20 data samples for the U.S. population precludes the latter determination. As a consequence we make the following stochastic assumptions:

The net reproduction rate, n(t), is a sample function of a random process characterized by:

A) for each t the samples n(0), n(1),...n(t) are identically distributed;
B) the random variables n(0), n(1),...,n(t) are statistically independent;
C) the ensemble average of n, E(n), exists for all t and E(n)>1;
D) the ensemble average of $\ell n(n)$ exists for all t and $E[\ell n(n)] < 0$.

Assumption C is motivated by the stationary maternity model. For that model the equilibrium $K_o(t) \equiv 0$ can be shown to be unstable, stable or asymptotically stable according as n >1, n=1 or n<1 respectively. Under assumption A and B the matrices A(0), A(1),...,A(t) are statistically independent and identically distributed and E(A) exists for all t. Thus

$$E[x(t)] = [E(A)]^t E[x(0)]. \tag{11}$$

It can be demonstrated that E(A) is a non-negative, irreducible and primitive matrix thereby having a real, positive, distinct and maximal characteristic value, λ, [3]. Therefore for sufficiently larger t we have the asymptotic condition

$$|E[x(t)]| \sim \lambda^t \tag{12}$$

From (12) we note that assumption C insures the expected equilibrium is unstable since C implies λ>1. As will now be shown assumption D guarantees that the sample equilibrium is asymptotically stable with probability 1.

THEOREM 1

Suppose assumptions A through D hold for the system x(t+1)=A(t)x(t). Then

$$\Pr\left[\lim_{t\to\infty} x(t) = 0\right] = 1.$$

Proof:

Under the norm assumed we have

$$|x(t+1)| \le |A(t)| \cdot |x(t)|$$

so that

$$\ell n|x(t+1)| \le \ell n|x(t)| + \ell n|A(t)|$$

which for any x(0) gives

$$|x(t)| \le |x(0)| \exp\left(\sum_{j=0}^{t-1} \ell n|A(j)|\right). \tag{13}$$

Since E(n)>1 there is a p>0 such that for all j

$$\Pr[\ell n|A(j)| = \ell n(n(j))] = p.$$

The sequence $\{\ell n|A(j)|\}$ j=0,1,... is a set of statistically independent, identically distributed random variables satisfying

$$E\left[\ell n|A(j)|\right]= p\ell n\left[n(j)\right].$$

By the strong law of large numbers the existence of $E[\ell n(n)]$ is necessary and sufficient for the truth of the following probability one statements:

1. $$\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} \left[\ell\ n|A(j)|-p\ell n(n(j))\right] = 0,$$

2. $$\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} \ell n(n(j))= E\left[\ell n(n)\right].$$

Statements 1 and 2 imply

3. $\sum_{j=0}^{t-1} \ell n|A(j)| \to pt\, E\left[\ell n(n)\right]$ as $t\to\infty$

Statement 3 and inequality (13) conclude the proof.

EXAMPLE:

To show that the theorem is not applicationally vacuous suppose n is distributed uniformly between 0.8 and 1.3. Then $E(n) = 1.05$; $E[\ell n(n)] = -0.078$ and $p = 0.6$. Thus for large t

$$|x(t)| \sim \exp(-0.0468t) \tag{14}$$

while

$$|E[x(t)]| \sim (1.05)^{t}. \tag{15}$$

Interestingly, estimate (15) predicts a doubling of the expected value every 15 time periods while (14) estimates a halving of the sample norm in the same time period.

COMMENTS ON THEOREM 1

The maternity matrix is a special case of a class of non-commuting, stochastic matrices. The problem of finding asymptotic expressions for the general case is a difficult one. The scalar case was solved first with results not fundamentally different from the above[4]. With considerably more assumptions on the A matrix asymptotically gaussian results can be obtained for the logarithm of $|A|$[5]. Moment calculations are facilitated by using Kronecker products or by taking advantage of the structure of $\underset{\sim}{A}$[6]. The general case can be solved in principle by transforming the x(t) process into a markov process on the unit sphere. The problem then becomes one of finding a probability measure for the transforming intermediate variable and this problem can in turn be expressed as an intractable integral equation[7]. Theorem 1 can be viewed as a direct consequence of the zero-one-law because of the statistical independence assumption on n(t)[8]. In this sense the proof only had to determine if convergence takes place with probability one or zero. Thus if the net reproduction rate is truly random and statistically independent then it is clear from the theorem that stabilizing the expected value of the birth rate, which must be done by making $E(n)=1$, can lead to extinction. The crucial question is whether or not n(t) is random and if random also statistically independent. It is important to note that n(t) reflects the behavior of the total reproductive population so that n(t) can meet the assumptions of the theorem even though, as probably is the case, the probability of a birth to a person in age class i+1 at time t depends on whether the person conceived while in age class i at time t-1.

STATISTICAL DEPENDENCE AND WEAK CONVERGENCE

Demographic data supports the structure of the population model developed but is insufficient in quantity and quality to support the statistical independence hypothesis on $n(t)$ so crucial to the strong convergence exhibited in Theorem 1. In this section we remove the independence assumption and prove a weak convergence theorem.

Let $n(t)$ be a sample function from a weakly ergodic random process and assume a joint probability density function exists for $n(j)$ and $n(j+1)$; $j=0,1,2,\dots,t$, satisfying for all j:

1. $E[\ln(n(j)] < 0$
2. $\Pr[n(j) > 1] = p$,
3. $\Pr[n(j) > 1; n(j+1) > 1] = q < p$
4. for positive constants $M_1 < M_2$, $\Pr[M_1 < n(j) < M_2] = 1$.

As in Theorem 1, assumption 1 is essential to the convergence results. Assumptions 2 and 3 insure that the possibility of birth rate growth is not precluded. Assumption 4 expresses the biological fact that there is an upper limit, M_2, to the numbers reproduced and that the probability of not a single birth occurring in a given 5 year period is 0.

Before proving Theorem 2 we indicate by an example that the above assumptions are not pathological in a demographic context. If one were to specify a joint density function for $n(j)$ and $n(j+1)$ with the aim of stabilizing the birth rate about some non-zero and finite value, one would probably be drawn to making the conditional probability specifications

$$\Pr[n(j+1) < 1 \mid n(j) > 1] \simeq \Pr[n(j+1) > 1 \mid n(j) < 1] \simeq 1$$

$$\Pr[n(j+1) > 1 \mid n(j) > 1] \simeq \Pr[n(j+1) < 1 \mid n(j) < 1] \simeq 0.$$

The first specification intimates that the reproductive population "looks back" 5 years (through the influence of the presently abundant or sparse child population) and curtails or increases the birth rate commensurate with birth rate stability. The second indicates that only under unusual circumstances would the reproducers act to deliberately send the birth rate to a projected equilibrium of 0 or ∞. Perhaps the simplest joint density function embodying these specifications is one uniform along the line

$$n(j+1) = -n(j)+2,$$

of width a, length b, and height $1/ab$. For $a=2.2$ and $b=0.5$ the parameters specified in assumptions 1 through 4 are

$$E[\ln(n(j))] = -0.142;\ p=0.5;\ q=0.057$$

$$M_1=1.954;\ M_2=0.046.$$

Additionally

$$\Pr[n(j+1)<1 \mid n(j)>1] = 0.886$$

$$\Pr[n(j+1)>1 \mid n(j)>1] = 0.114.$$

THEOREM 2

Suppose assumptions 1 through 4 hold for the system $x(t+1)=A(t)x(t)$. Then for any $\varepsilon>0$

$$\lim_{t\to\infty} \Pr[|x(t)| > \varepsilon] = 0.$$

PROOF:

In what follows all convergence statements are weak ones i.e. convergence in probability as contrasted with the almost everywhere convergence in Theorem 1. From previous work we have

$$|x(t)|\le|x(1)|\exp\left[\sum_{j=1}^{t}\ell n|A(j)|\right]. \tag{16}$$

From the weak law of large numbers we have that

$$\lim_{t\to\infty}\frac{1}{t^2}\operatorname{Var}\Big(\sum_{j=1}^{t}\ell n|A(j)|\Big)=0 \tag{17}$$

implies

$$\lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{t}\big[\ell n|A(j)|-E(\ell n|A(j)|)\big]=0.$$

Suppose (17) is satisfied. We know

$$E(\ell n|A(j)|)=p\,\ell n[n(j)]$$

and therefore by weak ergodicity

$$\sum_{j=1}^{t}\ell n|A(j)|\to pt\,E[\ell n(n(j))]. \tag{18}$$

(18), (16) and assumption 1 conclude the proof given (17) is true. To show (17) define $y(j)=\ell n|A(j)|$

$$s(t)=\sum_{j=1}^{t}y(j).$$

The pairwise dependence assumption on n(j) gives

$$\operatorname{Var}[s(t)]=\sum_{j=1}^{t}\operatorname{Var}[y(j)]+2\sum_{j=1}^{t}\big[E(y(j)y(j+1)-E(y(j))E(y(j+1))\big].$$

Assumptions 2 and 3 yield

$$E[y(j)y(j+1)]=q\,\ell n(n(j))\,\ell n(n(j+1))$$

$$E[y(j)]\;E[y(j+1)]=p^2\,\ell n(n(j))\,\ell n(n(j+1)).$$

Thus

$$\operatorname{Var}[s(t)]=(q-p^2)\sum_{j=1}^{t}\Big[\ell n^2(n(j))+2\ell n(n(j))\,\ell n(n(j+1))\Big].$$

By assumption 4 there is an $M_3>0$ such that

$$\operatorname{Var}[s(t)]\le 3t(q-p^2)\,\ell n^2(M_3)$$

ECONOMIC INTERACTIONS

The theorems derived would indicate that extinction is a clear possibility but the stochastic assumptions ignore significant influences economic variables have had on demographic parameters [9]. In this section we attempt to quantify with an expected value model one of the more recently proposed theories regarding economic and fertility behavior [10]. The bases of the theory or "Easterlin effect" are three unmistakable trends of social history. The first is the long term (since 1850) decline in the net maternity function accompanied by a long term increase in life expectancy. The second is the long term (since 1900) increase in the participation rate of women in the U.S. labor force. The third is the short term, (since 1940) rapid increase of younger age people attending college. The net effect of the second and third phenomena has been to depress reproduction in the younger age groups because of loss of income due to attending school while simultaneously younger and older females meet labor demand. Thus the second and third trends have, at least recently, precipitated the first.

In this section we will quantify this theory and indicate how it fits into the previous demographic model and into a neo-classical economic model. We take 30 year age classes as apposed to the previous 5 year groupings and time takes on 30 year increments. We assume mortality parameters are fixed. Thus equation (6) becomes

$$K_0(t) = g_1K_0(t-1) + g_2K_0(t-2) \tag{19}$$

and equations (1) becomes

$$K_1(t) = m_{1,0}K_0(t-1)$$

$$K_2(t) = m_{1,0}m_{2,1}K_0(t-2). \tag{20}$$

Equations (20) give the number of people between 0 and 30 and 31 and 60 respectively as a function of past birth rates and 30 year age grouping mortality parameters. The parameter g_1 in (19) is typically 0.8 of $g_1 + g_2$ and it is the one we subject to the Easterlin effect taking g_2 as constant. Central to this effect is the assumption that those in the early child-bearing ages derive their consumption behavior from their parents. If this is so then the desired consumption level inherited by the child bearers relates to the parents consumption level lagged one time period, i.e., at the time the child bearers were growing up under this last consumptive level. Thus we take as a measure of desired consumption in those $K_1(t)$ people the consumption level of those $K_2(t)$ people at time period t-1. We also assume that desired consumptive behavior is reflective of desired number of children. If $Y_1(t)$ and $Y_2(t)$ are dollar income to age groups 1 and 2 respectively then measures of per capita consumption are $Y_1(t)/K_1(t)$ and $Y_2(t)/K_2(t)$. The above argument says that $[Y_1(t)/K_1(t)]/[Y_2(t-1)/K_2(t-1)]$ is a measure of desired consumption for age class 1.

While age class 1 may have the desire for children it is also a fact that increasing labor force participation of women accompanies decreasing fertility. Let $f_1(t)$ be the percent of women in age class 1 participating in the U.S. labor force. Let $\hat{f}_1$ be the maximum expected value of $f_1(t)$. If the male experience is any indicator then $\hat{f} \cong 0.95$. Then we assume the Easterlin effect takes the quantitiative form

$$g_1(t) = g_1(t-1)\sigma_1(t)/\sigma_2(t-1) + b[\hat{f}-f_1(t)] \tag{21}$$

where $\sigma_i(t) = Y_i(t)/K_i(t)$ and b is a constant between 0 and 1. Recent regression results would indicate $b \cong 0.7$[11]. Equation (21) states that the child bearing age groups will reproduce as in the past if their consumption expectations are met ($\sigma_1 = \sigma_2$) and are unimpeded by the necessity of having the females work (f_1 small relative to $\hat{f}$).

The labor force, L(t), depends directly on the birth rate and labor participation rates. If $r_i(t)$, i=1,2,3 are the participation rates (male and female) of age classes 1,2 and 3 then

$$L(t) = r_i(t)m_{1,0}K_0(t-1) + r_2(t)m_{1,0}m_{2,1}K_0(t-2) \tag{22}$$
$$+ r_3(t)m_{1,0}m_{2,1}m_{3,2}K_0(t-3).$$

The connection between income, needed in (21), and labor force can be made through the neo-classical economic production function model. Let C(t) be net capital stock, Y(t) aggregate economic output, $\lambda(t)$ technological progress and s(t) the savings fraction of income. We also assume a constant elasticity of output relative to capital, α. Thus

$$\begin{aligned} Y(t) &= \lambda(t)\, L^{\alpha}(t)\, C^{1-\alpha}(t) \\ C(t+1) &= C(t) + s(t)Y(t) \end{aligned} \tag{23}$$

By defining $Z(t) = \frac{Y(t)}{C(t)}$ i.e. the output capital ratio, (23) can be put into the form of a first order nonlinear difference equation with three inputs, $\lambda(t)$, $L(t)$, and $s(t)$.

$$Z(t+1) = \left[Z(t)\ \lambda(t+1)/\lambda(t)\right]\left[\frac{L(t+1)/L(t)}{1+s(t)Z(t)}\right]^{\alpha}.$$

Finally we assume that age group income is proportional to the labor market share each age group enjoys. Hence

$$\begin{aligned} Y_1(t) &= \left[r_1(t)\ m_{1,0}K_0(t-1)/L(t)\right]Y(t) \\ Y_2(t) &= \left[r_2(t)\ m_{k,0}m_{2,1}K_0(t-2)/L(t)\right]Y(t). \end{aligned} \tag{25}$$

If we now consider equations (19),(21), (22), (24) and (25) we have a dynamic system in expected value variables with inputs $r_i(t)$, $i=1,2,3$, $s(t)$ and $\lambda(t)$. Note that all inputs reflect decision processes dependent on the future and seem natural choices for random variables. We investigate the asymptotic consequences of the birth rate reaching a non-zero expected value equilibrium. From (19) we must have $g_1 + g_2 = 1$ which implies from (21) that we must have $\sigma_1(t)/\sigma_2(t-1) \rightarrow \beta$ with $0<\beta\leq 1$. From the definition of $\sigma_1(t)$ we must have

$$\frac{r_1\ Y(t)}{L(t)} \rightarrow \beta\frac{r_2\ Y(t-1)}{L(t-1)} \tag{26}$$

where r_1 and r_2 are the asymptotic values of the labor force participation rates. In the economic sector the equilibrium of $Z(t)$, say Z^o, is, from (24),

$$Z^o(t) = \frac{[\lambda t+1)/\ \lambda(t)]^{\frac{1}{\alpha}}\ L(t+1)/L(t)-1}{s(t)}.$$

If Z^o is achieved and maintained then it is easily shown that in equilibrium

$$\frac{Y(t+1)}{Y(t)} = \left(\frac{\lambda(t+1)}{\lambda(t)}\right)^{\frac{1}{\alpha}}\ \frac{L(t+1)}{L(t)}\ . \tag{27}$$

From (26) and (27) we have (asymptotically)

$$\frac{\lambda(t+1)}{\lambda(t)} = \left(\beta\frac{r_2}{r_1}\right) \tag{28}$$

Equation (28) is intuitively appealing. If technological progress is to continue then r_2 must exceed r_1 asymptotically i.e. the older, more experienced worker must contribute more toward augmenting the labor supply. Also, obviously, $\lambda(t)$ and $r_i(t)$ will no longer be exogenous independently. It is also possible to show that if $Z^o(t)$ is maintained as an equilibrium, i.e., $Z^o(t)$ is a stable solution of (24) then $Z^o(t+1) = Z^o(t)$ from which we must have

$$\frac{Y(t+2)}{Y(t+1)} - \frac{s(t+1)}{s(t)}\ \frac{Y(t+1)}{Y(t)} + \frac{s(t+1)}{s(t)} - 1=0 \tag{29}$$

so that (29) and (27) imply $\lambda(t)$, $s(t)$ and $r_i(t)$ are not independent in equilibrium. Indeed, the $r_i(t)$ determine $\lambda(t)$ from (28), $L(t)$ from (22) and therefore, from (27) and (29), equilibrium savings ratio. As a consequence the labor force participation rates are, in equilibrium, the only exogenous variables. This is not an unexpected result for the Easterlin effect in essence assumes people trade off children and consumption and to consume one must work.

CONCLUSIONS

The population projection model of the U.S. used by the Census Bureau has a stochastically asymptotically stable null solution under assumptions that permit unbounded expected value growth in the population.

A model is proposed for controlling parameters of the projection matrix using economic variables. Under an assumption of expected value demographic and economic stability the controlling variables are labor force participation rates of the population in general and females in particular. It remains to be shown that the proposed model can be stabilized under conditions of stochastic disturbances.

BIBLIOGRAPHY

1. Keyfitz, N. Introduction to the Mathematics of Population, Addison Wesley, 1968.

2. Keyfitz, N. and Flieger, W. World Population) An Analysis of Vital Data, University of Chicago Press, 1968.

3. Gantmacher, F. R. The Theory of Matrices, Vol. 2, Chelsea, 1964.

4. Grenander, U. Probabilities on Algebraic Structures, Wiley, 1963.

5. Furstenburg, H. and Keston, H. "Products of Random Matrices," Ann. Math. Stat. Vol. 3, pp. 457-469, 1960.

6. Bellman, R. "Limit Theorems for Noncommutative Operations," Duke Math. J., Vol. 21, pp. 291-500, 1954.

7. Khas'Minskii, R. Z. "Necessary and Sufficeint Conditions for the Asymptotic Stability of Linear Stochastic Systems, "Theory of Prob. and Appl., Vol. 12, pp. 144-147, 1967.

8. Doob, J. L. Stochastic Processes, Wiley, 1953.

9. Keyfitz, N. Population Dynamics, Academic Press, pp. 1-38, 1972.

10. Easterlin, R. A., Population, Labor Force, and Long Swings in Economic Growth, Columbia U. Press, Chapters 5,6,7, 1968.

11. Gregory, P. et al. "A Simultaneous Equation Model of Birth Rates in the U.S.", Rev. Econ. and Stats., pp. 374-380, November 1972.

Subdiagonal of M		First Row of M	
Year=1920	1965	1920	1965
0.97351	0.99669	0.00000	0.00000
0.98854	0.99838	0.00067	0.00099
0.98610	0.99795	0.06783	0.08601
0.97653	0.99685	0.25360	0.32202
0.96943	0.99611	0.37263	0.43270
0.96640	0.99477	0.33123	0.31006
0.96432	0.99231	0.24978	0.17069
0.96141	0.98861	0.14663	0.07177
0.95495	0.98281	0.04922	0.01656
0.94187	0.97431	0.00706	0.00098

Table 1.

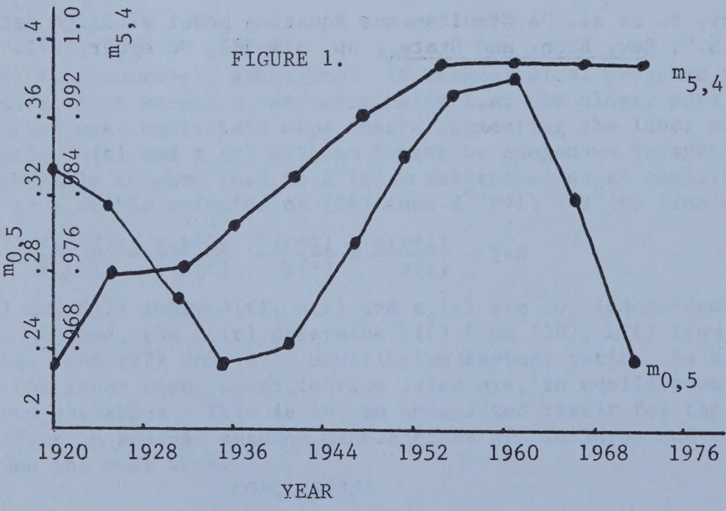


FIGURE 1.

INTERACTIONS BETWEEN POPULATION, PRODUCTION, AND MATERIAL BENEFIT SECTORS IN AN ECONOMETRIC MODEL OF JAPAN

KIMIO UNO
Japan Economic Research Center
Tokyo, Japan

ALAN I. DUCHAN
Canisius College
Buffalo, New York

ABSTRACT

This paper describes a portion of a recently developed econometric model of Japan. Emphasis in the model is on interactions between economic variables on the one hand and demographic and social variables on the other.

INTRODUCTION

Econometric models of national economies have usually treated population status variables as exogenously determined. Birth and death rates, age distribution variables, and the like affect the supply of labor and thus, the production sector; but there is no feedback from the latter sector to the population sector. Inasmuch as changes in production and material benefits affect demographic variables with a lag, the effect of such assumptions depends on the type of model. The longer the time horizon of the model, the greater will be the distorting effect of treating population variables as exogenous. Nonetheless, this assumption appears regularly, not only in short-run models where it may be reasonable, but in long-run growth models where it is dubious at best (see, e.g., [6]). The model presented here is a first step toward recognizing that, through its effect on the standard of living, production influences demographic variables.

The next section covers an overview of the model with an emphasis on the interactions between major sectors. Subsequent sections discuss material benefit, population size, and labor force participation sectors. A description of all variables appears at the end of the paper.

MODEL OVERVIEW

The work herein concerns a prototype econometric model aimed at revealing possible limits to economic growth, taking empirical data from the Japanese experience from 1950 to 1970. Among the model's innovative features is the inclusion of non-economic variables as an integral part. The model consists of twelve sectors, five of which involve more or less conventional economic variables. These sectors cover production, prices and wages, gross national product and gross national expenditures, capital formation, and foreign trade. The other sectors deal with social indicators, population size, labor force participation, environment, material benefits, density, energy and transportation, and speed of social transformation.

A complete discussion of the model appears in Uno's [7] monograph upon which this paper is based. We devote ourselves here to the material benefits, population size, and labor force participation sectors, introducing other sectors only as they interact with these three.

Among the variables determined in the production and economic sectors are production of consumer goods and social capital. Together, these determine the level of material benefits. In turn, per capita material benefits affect birth and death rates and life expectancy, thereby determining total population and the number of people over 15 years of age. The latter group is the working age population or potential labor force. In the labor force size sector, the potential labor force is reduced by the size of the student population with the remainder then divided into primary and secondary workers. Further reductions in the labor force are made for losses due to accidents and diseases, military service, imprisonment, labor disputes, and unemployment. These losses are determined by the per capita level of material benefits and, directly and indirectly, by the production sector. Also determined in the labor force size sector are hours worked. The number of employed workers and hours worked feedback to the economic sectors where they help determine the level of output and the wage level.

The parameter estimates shown below were obtained by ordinary least squares (OLS). As is well-known, OLS estimators of simultaneous equation parameters are inconsistent. However, the evidence regarding their small sample properties is mixed. In a number of simulation studies, OLS estimators were superior to conventional simultaneous equation estimators [3,5]. Furthermore, empirical work often produces essentially the same estimates for OLS and for simultaneous equation methods [1,2]. Because of these studies and because of the computational ease of obtaining OLS estimates, we decided to forgo the use of more sophisticated estimators. Yearly time series data were used, usually for the period 1949-70, but for a shorter period when data was unavailable. Data sources are given in [7].

MATERIAL BENEFITS SECTOR

This sector consists of one definitional equation, differing from most conventional measures of consumption in that it excludes consumer durables but includes estimates of the flow of services from durables and from other stocks. The measure of material benefits is:

$$MB = Cf + Cc + i \sum_{j=1}^{5} Cd_{t-j} + Kh + c \cdot Ks \ . \qquad (1)$$

The first two terms in (1) are consumption of food and of non-durables. The terms in brackets are estimates of stocks from which flow services to consumers. The first is an estimate of the stock of durable goods obtained by accumulating the flow of durables to consumers over the past five years. Theoretically, the accumulation should take place over the average life expectancy of the goods, but there is at present no way to estimate this amount. The next term is the stock of housing capital and the last is the stock of social capital (Ks) multiplied by an estimate of the share of social capital for final use (c).

To obtain the flow of services from the three stocks, an assumed user's benefit rate (i) of 10% was used. Ideally, this rate should be determined within a model involving the supply and demand for capital. However, the sample period was one of capital rationing so that interest rates did not reflect the true benefits of capital. Thus, empirical determination of the flow of capital services was virtually impossible.

POPULATION SIZE SECTOR

This sector consists of five regression equations and an identity:

$$\log b = -2.85 - 0.884 \log (MB/N)_{t-1} + 1.21\ D1 + 0.986\ D1 \cdot \log(MB/N)_{t-1} - 0.122\ D2 \qquad (2)$$

$$\log da = -2.66 - 0.412 \log(MB/N)_{t-1} + 0.406\ D1 + 0.332\ D1 \cdot \log(MB/N)_{t-1} \qquad (3)$$

$$\log db = -2.79 - 1.09 \log(MB/N)_{t-1} \qquad (4)$$

$$N_{t+1} = N[1 + b(1-db) - da + e] \qquad (5)$$

$$\log LE = 1.88 - 0.00336 \frac{1}{(MB/N)_{t-1}} \qquad (6)$$

$$\log(N^*/N) = -1.89 + 0.949 \log LE \qquad (7).$$

The main determinant of the birth and death rates (b, da, db) and of life expectancy (LE) is the level of per capita material benefits (MB/N), lagged one period. In addition, dummy variables were used to account for peculiarities within the sample period. Both the birth rate (b) and the non-infant death rate (da) experienced rapid declines from World War I to the late fifties. The dummy variable D1 (0 for 1950-57 and 1 for 1958-70) was used to account for this shift. Another dummy, D2, (1 in 1966 and 0 otherwise) was used to account for a dramatic drop in the birth rate in

1966, the year of "Hinoe-uma" or fire and horse. Women born in the year of Hinoe-uma are supposedly stronger than men and are said to eat men, making them undesirable mates. The sharpness and suddenness of the drop is remarkable. The birth rate in 1966 was 13.7 births per thousand population, 26% less than the 1965 rate and 29% less than the 1967 rate. One innovation of the present model is that death rates for infants (db) and non-infants (da) are separated. The advantage of so doing is that infant death rates are much more responsive to economic growth than are other death rates. No dummy was used in estimating db because it experienced a continued decline throughout the sample period with no leveling off after 1957.

Equation 5 determines population size. Except to note that the net immigration rate (e) is an exogenous variable in the model, (5) is self-explanatory.

In (6), per capita material benefits determine life expectancy (LE). From 1947 to 1971 LE has gone from 51.5 to 70.2 years for males and from 55.3 to 75.3 for females. As stressed by Kaneko [4], this marked improvement is due not only to improved current consumption, as conventionally measured, but to improved social facilities and other stocks, thus justifying our use of MB/N as the causal variable. A logarithmic transformation was used so that LE would have an asymptotic value rather than increasing indefinitely ($LE \rightarrow 10^{1.88}$ as $MB \rightarrow \infty$). Further, the appropriate functional form for LE should have LE = 0 when MB = 0. According to (6), LE is undefined when MB = 0, but since $LE \rightarrow 0$ as $MB \rightarrow 0$, we can define LE(0) = 0.

Finally, (7) determines the number of people over 15 (N*) as a fraction of N. LE is the causal variable since changes in LE change the age distribution of the population.

LABOR FORCE PARTICIPATION SECTOR

In this sector the working age population is reduced to account for persons that are in school, unemployed, or not working for a variety of other reasons. In addition, the labor force is divided into two parts--primary and non-primary. The sector consists of 9 regression equations and two identities:

$$\log(E/N^*_{t-1}) = -0.892 - 0.0129 \frac{1}{(MB/N)_{t-1}} \qquad (8)$$

$$\log(L/N^*) = 0.131 + 0.188 \log(E/N^*_{t-1}) + 0.433 \log(L/N^*)_{t-1} - 0.359\, D5 - 0.312\, D5 \cdot \log(E/N^*_{t-1}) \qquad (9)$$

$$\log(L2/L) = 0.0620 + 0.0157 \log(Y2/L2)_{t-1} + 0.854 \log(L2/L)_{t-1} \tag{10}$$

$$\log H = 2.36 + 0.0841 \log \frac{W2}{L2(1-u)} + 0.0596 \log(Q2_t/Q2_{t-1}) - 0.0960\ D4 - 0.138\ D4 \cdot \log \frac{W2}{L2(1-u)} \tag{11}$$

$$\log(1-u) = -0.0190 + 0.0281 \log(Q2_{t-1}/Q2_{t-2}) + 0.00647 \log Q2_{t-1} + 0.368 \log(1-u)_{t-1} \tag{12}$$

$$\log(1-ai) = -0.00117 + 0.0000450\ T \tag{13}$$

$$\log(1-at) = 0.000311 - 0.0000764 \log(TA \cdot wmv) \tag{14}$$

$$\log(1-ad) = 0.000720 + 0.00106 \log(MB/N)_{t-1} \tag{15}$$

$$\log(1-v) = -0.000338 + 0.000711 \log(MB/N)_{t-1} \tag{16}$$

$$L2^* = L2(1-u)(1-ai)(1-at)(1-ad)(1-v)(1-m)(1-\ell) \tag{17}$$

$$L1 = L - L2 \tag{18}$$

In (8), school attendance among the working age population was estimated as a function of per capita MB. Then in (9) the labor force participation rate is determined by its own lagged value and by E/N^*_{t-1}. D5 is a dummy variable (0 in 1950-54 and 1 in 1955-70) used to account for a shift in the function. The shift was due to a movement of the labor force from the primary sector, where the female participation has been high, to the non-primary sector, where it has been low.

In (10) the fraction of L in the non-primary sector is determined by its own lagged value and by value added per worker in the non-primary sector (Y2/L2). Value added in the non-primary sector serves as a proxy for income inducements and whatever other mechanisms attract the labor force out of the primary sector.

Next, hours worked was estimated as a function of the wage rate and the phase of the business cycle as measured by $Q2_t/Q2_{t-1}$. A dummy variable was used to distinguish two periods: 1950-60 and 1961-70. The dummy was needed because hours worked in the earlier period was limited by the effects of World War II on the level of the capital stock.

In (12) through (16), labor force reduction factors to account for unemployment, diseases and accidents, and imprisonment were estimated. The employment rate in the non-primary sector (1-u) is a function of its lagged value, cyclical fluctuations as measured by $Q2_{t-1}/Q2_{t-2}$, and growth in the economy as measured by $Q2_{t-1}$. Industrial accidents (ai) is determined by a time trend which serves as a proxy for improvements in safety measures. Traffic accidents (at) are explained by the level of motor vehicle activities (TA•wmv). The levels of infectious diseases (ad) and imprisonment (v) are explained by the level of per capita material benefits.

Finally, the effective non-primary labor force ($L2^*$) is defined. This is the labor force net of the adjustments given by (12-16) and of losses due to military service (m) and labor disputes (ℓ), both of which are treated as exogenous. We assumed that unemployment, diseases, etc. are independent of each other, allowing the use of a simple multiplicative form for (17). As a last step, the primary labor force (L1) is defined as a residual. As regards feedback to other sectors, both hours worked and the size of the primary and non-primary labor forces affect output, prices and wages.

CONCLUDING REMARKS

Simulations covering the sample period closely mirrored the observed time paths, as usually happens with time series data. As data for years outside the sample period become available, we hope to test the models forecasting ability. Meanwhile, the model requires a number of refinements. In particular, an important part of the model that has not been discussed here is the environment sector. In this sector the levels of various productive activities determine the levels of a number of pollutants. In future work, we hope to have this sector feedback to affect production, investment in pollution control capital, and labor force losses due to diseases.

LIST OF VARIABLES

ad	Ratio of man-years lost due to infectious diseases to total man-years of labor.
ai	Ratio of man-years lost due to industrial accidents to total man-years of labor.
at	Ratio of man-years lost due to traffic accidents to total man-years of labor.
b	Birth rate.
Cc	Consumption of non-durable consumer goods in 1965 prices (billion yen).
Cd	Consumption of durable consumer goods in 1965 prices (billion yen).
Cf	Consumption of food in 1965 prices (billion yen).
c	Share of social capital for final use.
D1	Dummy variable: 0 for 1950-57 and 1 for 1958-1970.
D2	Dummy variable: 1 for 1966 and 0 otherwise.
D4	Dummy variable: 0 for 1950-1960 and 1 for 1961-1970.
D5	Dummy variable: 0 for 1950-1954 and 1 for 1955-1970.
da	Death rate for non-infants.
db	Death rate for infants (per live birth).
E	Number of people 15 years of age and older attending educational institutions (thousands).
e	Net immigration rate.
H	Work hours per man per month in non-primary sector.
i	User's benefit rate for final use capital goods.
Kh	End of year housing capital stock in 1965 prices (billion yen).
Ks	End of year social capital stock in 1965 prices (billion yen).
L	Total labor force (thousands).

L1	Primary sector labor forces (thousands).
L2	Non-primary sector labor force (thousands).
L2*	Non-primary sector labor force net of unemployment, industrial accidents, traffic accidents, infectious diseases, imprisonment, military personnel, labor disputes (thousands).
LE	Life expectancy, average of male and female (years).
ℓ	Ratio of man-years lost due to labor disputes to total man-years of labor.
MB	Material benefits in 1965 prices (billion yen).
m	Ratio of military personnel to non-primary labor force.
N	Total population (thousands).
N*	Population 15 years of age or older (thousands).
Q2	Index of non-primary sector output (1965 = 100).
TA	Transportation activities, total of freight ton-km and passenger-km (millions).
T	Time trend (= 1 in 1950).
u	Unemployment rate in the non-primary sector.
v	Imprisonment rate among persons 15 years of age or older.
W2	Compensation of employees in the non-primary sector in current prices (billion yen).
wmv	Share of motor vehicles in transportation activities.
Y2	Index of value added in the non-primary sector in current prices (billion yen).

REFERENCES

[1] Christ, C. F.: *Econometric Models and Methods*. New York: John Wiley & Sons, 1966.

[2] Fox, K. A.: "Econometric Models of the United States," *Journal of Political Economy*, 64 (1956), 128-142.

[3] Johnston, J.: *Econometric Methods*, 2nd ed. New York: McGraw-Hill, 1972.

[4] Kaneko, T.: "Shogaikokutono Hikakukara Mita Nihonno Heikin Yomei (Recent Trends in the Life Expectancy in Japan and Some International Comparisons)," *Jinko Mondai Kenkyujo Nenpo* (Annual Reports of the Institute of Population Problems), 1971, Tokyo (1971), 67-72.

[5] Malinvaud, E.: *Statistical Methods of Econometrics*, 2nd revised ed. Amsterdam: North-Holland, 1970.

[6] Morishima, M., *et al.*: *The Working of Econometric Models*. Cambridge: Cambridge University Press, 1972.

[7] Uno, K.: "The Limits of Japanese Growth: An Econometric Study," Unpublished Ph. D. dissertation, University of Illinois at Urbana-Champaign, 1972.

AN APPLICATION OF THE LINEAR-QUADRATIC REGULATOR PROBLEM TO AN ECONOMIC SYSTEM

Kamran Amir-Ebrahimy and William G. Vogt
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, Pennsylvania 15261

ABSTRACT

The problem of the determination of optimal decision rules concerning the dynamic transition of a firm facing a monopolistic market from a given initial condition to a known statically optimal equilibrium condition is formulated and solved as a linear quadratic regulator problem. The solution, in general, is shown to consist of two parts: a maximum effort portion followed by a singular control.

INTRODUCTION

Mathematical modeling of business and economic activities has attracted people of diverse backgrounds for quite some time. The recent literature on the theory of the firm demonstrates a marked tendency on the part of some economists to move from neoclassical models to the so called behavioral or managerial models.[1] Much of the available literature regarding the subject deals with operations research type development of optimal inventories based on generally fixed prices and given production schedules.[2,3,4] The industrial dynamics literature[5] on the subject, though recognizing the interdependence of such variables as production, prices, sales, etc., concentrate primarily on the modeling aspects of the problem. The "correct" decision could presumably be deduced from comparing different simulation runs of the model. With the notable exception of George Hay's article, few of the recent works have adopted a systems viewpoint with price and production as control variables, where decision making is viewed as a "single optimization process from which all relevant variables result."[6]

The objective of this paper is to investigate optimal decision rules on pricing, production, and inventories based on a behaviorial model of a firm facing a monopolistic market. It is common knowledge among economists that, in a monopoly, profits are maximized when marginal cost, MC, equals marginal revenue, MR. The optimum point of operation for the firm depends, among other things, on the production level of the firm as well as the market demand. Clearly, this point shifts, should there be a significant change in the parameters related to market demand or production.

Large industrial producers and manufacturers - for instance, auto manufacturers, steel or aluminum producers, etc. - despite the generally held view of operating in an oligopolistic market, are appropriate examples, provided there exists enough product differentiation so that the firm has some control over the price. In other words, a key assumption is that the firm in question may be viewed as an isolated unit in the market with some degree of independence as far as price setting is concerned.

The dynamics of the system will be analyzed in a discrete time context. The number of the new orders for goods the firm receives is taken as a proxy for the actual demand for goods. It is clear that in a situation where the number of orders placed on the firm over a certain period exceeds (or falls short of) the production, an unbalance exists which cannot continue for ever as it will ultimately lead to the depletion of the inventories (or overstocking). The point of interest is that the excess demand may be countered by either an increase (decrease) in the production rate, or increase (decrease) in prices, or a combination of both. This paper addresses itself to the determination of optimal policies governing the dynamic transition of the operating state of the firm from a given set of initial conditions to a final "equilibrium" state, which is determined by the rule of static optimality.

$$MR = MC \tag{1}$$

TERMINOLOGY AND SYSTEMS DYNAMICS

The following notation is used to identify the variables of interest:

P_k	production in period k; number of units of goods produced in period k
I_k	inventory in period k; if positive, (available for sale) units of goods passed on to next period; if negative, the absolute value represents the number of unfilled orders in period k
q_k	market demand in period k; taken as the number of orders received in period k
O_k	number of orders due for shipment in period k
p_k	price per unit output in period k (a control variable)
ΔP_k	increase in the production in period k (a control variable)

System dynamics are assumed to be governed by a few relatively simple equations

$$P_{k+1} = P_k + \Delta P_k \tag{2}$$

$$I_{k+1} + O_{k+1} = I_k + P_{k+1} \tag{3}$$

$$q_{k+1} = a_o q_k + a_1 q_{k-1} + \dots + a_{n-3}\, q_{k-n+3} + a + b p_k \tag{4}$$

$$O_k = q_{k-1}{}^{*} \tag{5}$$

---

[*] Note that although O_k is assumed to be equal to the market demand delayed by one period, we may allow any number of delays with no significant change in the results.

Equation (4) represents a dynamic partial adjustment demand model. b is negative and indicates the short term relation between demand and price. a is positive and accounts for exogenous effects - such as the income variable, or seasonal changes - influencing the market. $a_0, a_1, \ldots, a_{n-3}$, $(a_{n-3} \neq 0)$, are adjustment parameters, constrained by the stability conditions for the model. All of these parameters are assumed constant and known in the present model. This implies short run demand curve shifting up and down as the effective value of a varies from one period to the next (see Figure 1).

Equations (2-5) can be handled much more efficiently by letting

$$
\begin{aligned}
x_k^1 &\triangleq P_k & \qquad u_k^1 &\triangleq \Delta P_k \\
x_k^2 &\triangleq I_k & \qquad u_k^2 &\triangleq p_k \\
x_k^3 &\triangleq q_k \quad ; \\
x_k^4 &\triangleq q_{k-1} = 0_k \\
x_k^5 &\triangleq q_{k-2} \\
&\vdots \\
x_k^n &\triangleq q_{k-(n-3)}
\end{aligned}
\tag{6}
$$

and,

$$\underline{X}_k \triangleq (x_k^1, x_k^2, \ldots, x_k^n)^T \quad , \quad \underline{U}_k \triangleq (u_k^1, u_k^2)^T \tag{7}$$

Now the dynamics of the transition of the firm may be represented by the following linear, time-invariant, discrete system

$$\underline{X}_{k+1} = \underline{\Phi}\,\underline{X}_k + \underline{B}\,\underline{U}_k + \underline{C} \qquad k = 0,1,2,\ldots \tag{8}$$

$\underline{\Phi}$, $\underline{B}$, and $\underline{C}$ have the form

$$
\underline{\Phi} = \begin{bmatrix}
1 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\
1 & 1 & -1 & 0 & \cdot & \cdot & \cdot & 0 \\
0 & 0 & a_0 & a_1 & \cdot & \cdot & \cdot & a_{n-3} \\
\cdot & \cdot & 1 & 0 & \cdot & \cdot & \cdot & 0 \\
\cdot & \cdot & 0 & 1 & & & & \cdot \\
\cdot & \cdot & \cdot & \cdot & & & & \cdot \\
\cdot & \cdot & \cdot & & \cdot & & & \cdot \\
\cdot & \cdot & \cdot & & & \cdot & & \cdot \\
0 & 0 & 0 & & & & 1 & 0
\end{bmatrix}
, \quad
\underline{B} = \begin{bmatrix}
1 & 0 \\
1 & 0 \\
0 & b \\
0 & 0 \\
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot \\
\cdot & \cdot \\
0 & 0
\end{bmatrix}
, \quad
\underline{C} = \begin{bmatrix}
0 \\ 0 \\ a \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0
\end{bmatrix}
\tag{9}
$$

The system $\underline{\Phi}$, $\underline{B}$, $\underline{C}$ is controllable.

It can be shown that there exists an equilibrium state $\underline{X}_e$ and an equilibrium control $\underline{U}_e$ with the property that once the system is at the equilibrium state, it can be maintained there by the application of the equilibrium control.

$$\underline{X}_e = \underline{\Phi}\,\underline{X}_e + \underline{B}\,\underline{U}_e + \underline{C} \tag{10}$$

The (long-term) equilibrium conditions may be determined by the application of the rule of static optimality (1) to the long-term demand equation

$$q_\ell = \frac{a}{1-\sum\limits_{i=0}^{n-3} a_i} + \frac{b}{1-\sum\limits_{i=0}^{n-3} a_i}\,p \tag{11}$$

for which marginal revenue is given[8] by the linear equation

$$MR = \frac{a}{1-\sum\limits_{i=0}^{n-3} a_i} + \frac{2\,b}{1-\sum\limits_{i=0}^{n-3} a_i}\,p \tag{12}$$

Hence,

$$\underline{X}_e = [q_e, q_e, \ldots, q_e]^T \quad , \quad \underline{U}_e = [0,\ p_e]^T$$

where p_e is the equilibrium price, and q_e the equilibrium demand (see Figure 2).

Since it is known that optimal operation results when the system is at the equilibrium state, the criterion of optimality is chosen to be the minimization of the deviations of the state of the system from the equilibrium values - in a least square error sense - over the entire transition period. Intuitively, this minimization is analogous to requiring that the system get close to the equilibrium point "as fast as possible." Furthermore, in recognition of real world constraints associated with price and production, we will assume upper and lower bounds on the components of the control such that

$$\begin{aligned} &\Delta P_{min} < 0 < \Delta P_{max} \\ &p_{min} < p_e < p_{max} \end{aligned} \tag{14}$$

The set of admissible controls will be denoted by $\mathcal{U}$

$$\mathcal{U} = \{\underline{U} \in R_2 \mid U^i_{min} \leq U^i \leq U^i_{max} \quad , \quad i=1,2\} \tag{15}$$

The control effort is assumed zero-cost in this paper. The case of nonzero-cost controls is considered in a separate paper to appear.

PROBLEM FORMULATION

The problem is tantamount to choosing the sequence $\underline{U}_0, \underline{U}_1, \underline{U}_2, \ldots,$ so as to minimize the cost functional

$$J = 1/2 \sum_{k=0}^{\infty} \| \underline{X}_k - \underline{X}_e \|^2_{\underline{Q}} \tag{16}$$

subject to (8) and to the additional conditions

$$\underline{X}(0) = \underline{X}_0 \; ; \; \underline{X}_\infty = \underline{X}_e \; ; \; \underline{U}_k \in \mathcal{U} \quad , \quad k = 0,1,2,\ldots \tag{17}$$

where $\underline{Q}$, the constant cost matrix, is an nxn positive (semi-) definite matrix. Two reasonable choices of $\underline{Q}$ are

$$\underline{Q}_1 = \begin{bmatrix} q_{11} & & & \\ & q_{22} & & \underline{0} \\ & & \ddots & \\ & \underline{0} & & q_{nn} \end{bmatrix} \; ; \; \underline{Q}_2 = \begin{bmatrix} \alpha & & & & & \\ & \frac{\beta}{2} & & \frac{-\beta}{2} & & \\ & & \frac{\gamma}{n-2} & & \underline{0} & \\ & \frac{-\beta}{2} & & \frac{\gamma}{n-2} + \frac{\beta}{2} & & \\ & & & & \frac{\gamma}{n-2} & \\ & & \underline{0} & & & \ddots \\ & & & & & \frac{\gamma}{n-2} \end{bmatrix} \tag{18}$$

$\underline{Q}_2$ being inspired by a performance index such as

$$J = 1/2 \sum_{k=0}^{\infty} \{\alpha(P_k - P_e)^2 + \beta(I_k - O_k)^2 + \gamma\,(q_k - q_e)^2\} \tag{19}$$

which has obvious implications.

PROBLEM SOLUTION

To solve the discrete linear-quadratic problem formulated above, first note that the discrete maximum principle applies in this case[9]. The maximum principle conditions may be determined by defining the Hamiltonian for the system and setting up the canonical equations of motion, along with the appropriate boundary conditions.

$$H(\underline{X}_k, \underline{\lambda}_{k+1}, \underline{U}_k) = \frac{1}{2} \left|\left|\underline{X}_k - \underline{X}_e\right|\right|^2_{\underline{Q}} + \underline{\lambda}^T_{k+1} \left(\underline{\Phi}\underline{X}_k + \underline{B}\,\underline{U}_k + \underline{C}\right) \qquad (20)$$

$$\underline{\lambda}_k = \underline{Q}\,(\underline{X}_k - \underline{X}_e) + \underline{\Phi}^T\,\underline{\lambda}_{k+1} \qquad (21)$$

$$\underline{X}_{k+1} = \underline{\Phi}\,\underline{X}_k + \underline{B}\,\underline{U}_k + \underline{C} \quad , \quad k = 0,1,2,\ldots \qquad (22)$$

$$\underline{X}(0) = \underline{X}_0 \quad , \quad \underline{X}_\infty = \underline{X}_e$$

Accordingly, if $\underline{U}_k^*$ is the optimal control,

$$H(\underline{X}_k, \underline{\lambda}_{k+1}, \underline{U}_k^*) = \min_{\underline{U}_k \varepsilon \mathcal{U}} H(\underline{X}_k, \underline{\lambda}_{k+1}, \underline{U}_k) \qquad (23)$$

So that

$$U_k^{i*} = \begin{cases} U^i_{\max} & \text{if} \quad \underline{\lambda}^T_{k+1}\,\underline{B}_i < 0 \qquad (24a) \\ U^i_{\min} & \quad\;\; \underline{\lambda}^T_{k+1}\,\underline{B}_i > 0 \qquad (24b) \\ U^i_{\min} \leq U_k^{i*} \leq U^i_{\max} & \quad\;\; \underline{\lambda}^T_{k+1}\,\underline{B}_i = 0 \qquad (24c) \end{cases}$$

where $\underline{B}_i$, i=1,2, is ith column of $\underline{B}$

$$\underline{B} \triangleq (\underline{B}_1 \,\vdots\, \underline{B}_2) \qquad (25)$$

As the Hamiltonian is linear in the control for this system, and each component of control is bounded, the optimal controller must be maximum effort (or singular, if a singular solution is admissible). Furthermore, in this case the maximum principle conditions are both necessary and sufficient for optimality[10].

Equation (24c) corresponds to the singular case where H does not depend on the ith component of the control vector. For the system discussed above, it is generally possible to have a totally singular solution

$$\underline{B}^T\,\underline{\lambda}_{k+1} = \underline{0} \qquad (26)$$

To show this, let

$$\underline{\lambda}_k = \underline{K}_k \underline{X}_k + \underline{G}_k \tag{27}$$

From (21), (22), and (26), it follows that the singular control is obtained by

$$\underline{U}_k = -(\underline{B}^T \underline{K}_{k+1} \underline{B})^{-1} \underline{B}^T (\underline{K}_{k+1} \underline{\Phi} \underline{X}_k + \underline{K}_{k+1} \underline{C} + \underline{G}_{k+1}) \tag{28}$$

where

$$\underline{K}_k = \underline{Q} + \underline{\Phi}^T \underline{K}_{k+1} \underline{\Phi} - \underline{\Phi}^T \underline{K}_{k+1} \underline{B} (\underline{B}^T \underline{K}_{k+1} \underline{B})^{-1} \underline{B}^T \underline{K}_{k+1} \underline{\Phi} \tag{29}$$

$$\underline{G}_k = -\underline{Q}\, \underline{X}_e + \underline{\Phi}^T [\underline{I} - \underline{K}_{k+1} \underline{B} (\underline{B}^T \underline{K}_{k+1} \underline{B})^{-1} \underline{B}^T] (\underline{K}_{k+1} \underline{C} + \underline{G}_{k+1}) \tag{30}$$

It turns out there exist constant matrices $\underline{K}$ and $\underline{G}$ satisfying Equations (28-30), such that

$$\underline{G} = -\underline{K}\, \underline{X}_e \tag{31}$$

$$\underline{U}_k = \underline{U}_e - (\underline{B}^T \underline{K}\, \underline{B})^{-1} \underline{B}^T \underline{K}\, \underline{\Phi} (\underline{X}_k - \underline{X}_e) \tag{32}$$

The singular region will then be the part of the intersection of two hyperplanes in the n-dimensional state space, where the singular control obtained by (32) satisfies the control constraints. Thus, the totally singular region is the (n-2)-dimensional subspace S_T

$$S_T = \{\underline{X} \,\varepsilon\, R_n \mid \underline{B}^T \underline{P} (\underline{X} - \underline{X}_e) = \underline{0},\ \underline{U}_e - (\underline{B}^T \underline{K}\, \underline{B})^{-1} \underline{B}^T \underline{K}\, \underline{\Phi} (\underline{X} - \underline{X}_e) \,\varepsilon\, \mathcal{U}\}$$

Obviously $\underline{X}_e \,\varepsilon\, S_T$. It can also be shown that once the system is in the singular region, the optimal strategy is to maintain the system in the singular mode by the application of a sequence of singular controls to ultimately drive the system to the equilibrium point. Furthermore, the final portion of all optimal arcs reaching the equilibrium point is wholly within this singular region[11].

In general, the optimal strategy consists of a sequence of maximum effort controls followed by a set of singular controls. Details of a quantitative nature, such as the precise duration of the maximum effort portions, or the value of the singular control are the subject matter of a numerical solution and potentially involve certain numerical complications. Such details, which clearly depend on the specific parameters and the initial conditions of the problem, are beyond the scope of this paper and will be reported in a later work.

CONCLUSION

This paper presents a new approach to the mathematical modeling and optimization of a firm facing a dynamic monopolistic market. The proposed model is linear, discrete time, and behavioristic in nature, leaving ample space for modifications and further extensions. The criterion of optimality employed is getting the system to coast off to the "equilibrium point" while certain desired criteria are achieved as best as possible. A salient feature of the system is that the terminal portion of every optimal trajectory transfering the system to the desirable operating point is a singular arc. The optimal strategy, in general, consists of the application of a bang-bang control to transfer the system to the singular region followed by a totally singular portion.

REFERENCES

1. R.M. Cyert and C.L. Hedrick, "Theory of the Firm: Past, Present and Future; An Interpretation," Journal of Economic Literature, June 1972, pp. 398-412.
2. K.J. Arrow and S. Karlin and H. Scarf (eds), Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, 1958.
3. G. Hadley and T.M. Whitin, Analysis of Inventory Systems, Prentice Hall, 1963.
4. H.A. Taha, Operations Research, An Introduction, MacMillan, 1971.
5. J.W. Forrester, Industrial Dynamics, MIT Press, 1961.
6. G.A. Hay, "Production, Price and Inventory Theory," American Economic Review, Sept. 1970, pp. 531-545.
7. H.S. Houthakker and L.D. Taylor, Consumer Demand in the United States: Analysis and Projections, (2nd and enlarged edition), Harvard University Press, 1970.
8. F.M. Scherer, Industrial Market Structure and Economic Performance, Rand McNally and Company, 1970.
9. J.M. Holzman, "Convexity and the Maximum Principle for Discrete Systems," IEEE Trans. on Aut. Cont., Vol. AC-11, No. 1, Jan. 1966.
10. L.I. Rozonoer, "L.S. Pontryagin's Maximum Principle in Optimal Control Theory," Automat. i. Telemekh., 20 (1959): 1320-34, 1441-58, 1561-78. Translated in Automation and Remote Control, 20 (1960): 1288-302, 1405-21, 1517-32. Reprinted in R. Oldenburger (ed), Optimal and Self-Optimizing Control, MIT Press, 1966.
11. R.E. Kalman, "On the General Theory of Control Systems," Proc. 1st Intr. Congr. Aut. Cont., London, Eyre & Spottiswood, 1961.
12. D.H. Jacobson, "Totally Singular Quadratic Minimization Problems," IEEE Trans. on Aut. Cont., Vol. AC-16, No. 6, Dec. 1971.

13. J.W. Graham and A.F. D'Souza, "Singular Optimal Control of Discrete Systems," 1970 Joint Automatic Control Conference, Preprints.

14. T.J. Tarn, S.K. Rao, J. Zaborszky, "Singular Control of Linear Discrete Systems," IEEE Trans. on Aut. Cont., Vol. AC-16, No. 5, Oct. 1971.

15. J. Zaborszky and T.J. Tarn, "Further Results on the Singular Control of Discrete Systems," IEEE Trans. on Aut. Control, April 1972.

16. P. Dorato, A.H. Levis, "Optimal Linear Regulators, The Discrete-Time Case," IEEE Trans. on Aut. Control, Vol. AC-16, No. 6, Dec. 1971.

17. B.D.O. Anderson, "Partially Singular Linear-Quadratic Control Problems," IEEE Trans. on Aut. Control, Aug. 1973.

18. C.T. Chen, Introduction to Linear Systems, Holt, Rinehart and Winston, Inc., 1970.

19. A.P. Sage, Optimum Systems Control, Prentice Hall, 1968.

20. A.E. Bryson, Jr. and Y.C. Ho, Applied Optimal Control, Blaisdell Publishing Co., 1969.

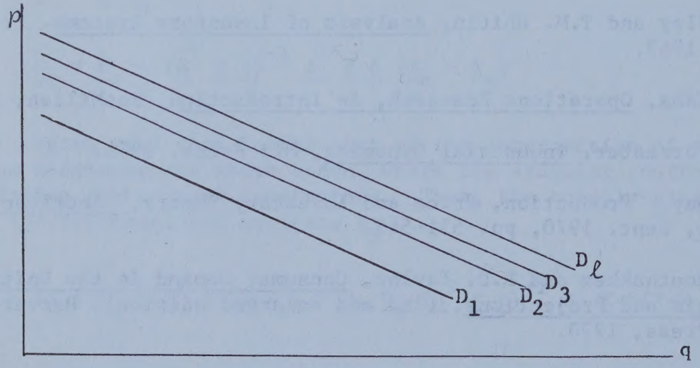


Figure 1. Short Run Demand Curve Variation

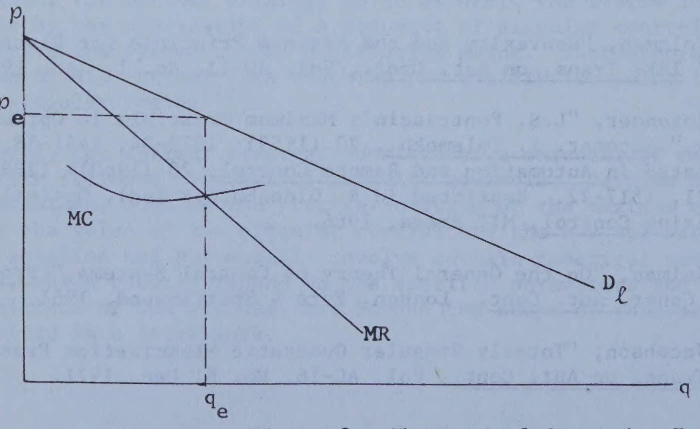


Figure 2. The Optimal Operating Point

CALIBRATION OF AN URBAN SIMULATION MODEL ACROSS CITIES

Richard B. Parker*
Economics Department, University of Illinois, Urbana, Il. 61801

ABSTRACT

Calibration problems arising from (1) location-specific model structure and (2) data deficiencies are analyzed. Examples drawn from a calibration experiment with the NBER Urban Simulation Model suggest some design guidelines for broadening model application with minimal loss of precision.

I. INTRODUCTION

History

Among economists interested in computer simulation methods, there has been in recent years a rapid growth of interest in their application to the modeling of urban economic phenomena, especially the long-run dynamics of housing, transportation, and land use.[1] While earlier efforts to construct large-scale econometric models of the aggregate national economy may have fallen short of expectations regarding their usefulness for policy purposes, the urban simulation models have not yet reached that full stage of development at which a definitive evaluation could be made. As there has been sharp ctiticism from some quarters of past efforts [4] and even of the basic simulation approach [1], an initial word or two in defense of the methodology seems in order, along with a suggestion of its broad applications.

Why Urban Simulation Models?

Structural/theoretical studies. Formal analysis of an economic system, in which a model's dynamic properties are inferred mathematically from its structure, is severely limited by size and linearity constraints. The complexity and diversity of economic phenomena is so great that in most cases of interest the links between structure and behavior cannot be directly exposed by formal analysis, so that we must instead grope for understanding through artificial simulation. Thus, if one could construct a sufficiently general model of urban housing markets, simulation experiments might disclose some rather general structural relationships among racial discrimination, housing prices, land use, and neighborhood quality, or between commuting cost and household location.

Forecasting/experimental applications. In addition to providing new theoretical insights, urban simulation models have potentially wide application in forecasting and policy experiments by planners and policy analysts. The long-run changes in urban structure induced by alternative routes for a new transport artery could thus be assessed in advance, as could changes in zoning, tax policies, or public facility location.

---

* The author is grateful to Royce Ginn, William Apgar, and the National Bureau of Economic Research for help in using their model; and to Noel Uri, William Hausman, Prentice Knight, and the other students in his graduate seminar, who did most of the work.

[1] See Nagelberg [6] for a selected bibliography. Mills [5] provides a synopsis and evaluation of several major efforts by economists.

The Dilemma of Local Calibration

Land use, demographic characteristics, income and employment structure, distribution of housing types, and many other dimensions of urban structure exhibit substantial variance across cities within the United States. In all but the simplest of urban models, this variance is sufficient to require city-by-city estimation of at least some of the key parameters. This local "calibration" of an urban model is clearly a necessary condition for its use in policy experiments. As the dimensions and complexity of an urban model increase, so do the difficulties of local calibration, until a level is reached at which the model itself contains structural characteristics or data requirements unique to a particular city. In this way, the research economies of scale realized by a more general model with broad applicability are sacrificed for greater precision in a particular application. This key dilemma facing the builder of urban models is the subject of this paper.

II. ISSUES IN MODEL DESIGN AND CALIBRATION

Research in the specification and estimation of econometric models has generated since the late 1930's a vast literature.[2] This paper is concerned with two dimensions of the problem which are of particular importance in the modeling of urban economic phenomena: the tendency to include structural elements which are unique to the specific city being modeled, and the related tendency to define variables so as to match available data, which may also be location-specific. Some generalizations can be made, before we look at an illustration.

Structural Problems

Models of urban structure must of course, be estimated using data from real cities. Even when the objective is to construct a general urban model with broad applicability, one must rely on empirical studies which typically involve only a single city. When specifying equations to explain, for example, the demand for housing, the question of whether to include trailers as a housing type may hinge on location-specific characteristics: e.g., whether trailers are a feasible housing type in the city under study. Thus, a housing-sector model derived from a study of a city where trailers were not a significant housing type may not form an adequate basis for a more general model. While it may often be possible to include in a general model a variable omitted from a specific study, the converse case also arises and creates further difficulties. The temptation to add variables in order to increase the explanatory power of the model is very strong; some of these will be unique to the city under study, and the model structure will thus no longer be generalizable. The ultimate answer to the question, "How general can we make the specification of an urban simulation model?" must always be determined by the generality of its underlying behavioral relationships. Ideally, the general model should be based on empirical studies of its component sectors using data from a diverse cross-section of cities. Such studies are still relatively scarce, in part due to the data problems which we will consider next. Even if they were plentiful, we could not fully escape the dilemma created by sector interdependence: the underlying behavioral relationships which we need to know in advance are part of what the general simulation model seeks to discover.

---

[2] For a recent review of this literature, see Judge [3].

Data Problems

More immediate obstacles to the construction of general urban models and to their calibration across cities are created by the data. Within the boundaries of data generated by a single agency such as the U.S. Census, things are not so bad; while there is some variation in the detail and disaggregation of Census data according to city size or SMSA status, the variables are in general defined and measured with reasonable consistency for a large set of cities. The problem, of course, is that we are usually not content to remain within these data boundaries. If theory suggests that the journey-to-work is an important determinant of household location, our model will require data on the workplace locations of households, which comes not from the Census but from origin-destination or household interview surveys typically undertaken on an individual SMSA basis as part of a transportation study tailor-made for each metropolitan area. The result is a splendid patchwork of data presenting a formidable challenge to the investigator who would like it to be consistent and commensurate across even a small cross-section of cities. Thus, the great majority of urban simulation models designed by economists to date have been models of a single metropolitan area. In most cases, data availability has been a main determinant of model structure.

III. AN ILLUSTRATION: THE NBER MODEL

"Pittsburgh II" is the name of the current version of an urban simulation model which has been under development for about a decade. Since 1968, the work has been under the aegis of the National Bureau of Economic Research (NBER). This section contains a brief description of the NBER model and a report of the problems encountered in the author's calibration experiment using Indianapolis data. For a full description of an earlier version of the model, the Detroit prototype, see Ingram [2].

Structure of the NBER Model

While the NBER Urban Simulation Model includes such variables as changes in industrial location and in urban demographic structure, its central concern is with the modeling of urban housing markets, that is, with the behavior over time and space of housing consumers, housing suppliers, and the clearing of the resulting housing "market." The seven submodels can thus be viewed as carrying out the activities of three main economic sectors: a housing demand sector, a housing supply sector, and a market clearing sector. Beginning with exogenous changes in industry employment, the demand sector represents changes in the level and mix of employment at each workplace, the decisions of households to move, and the choice of housing "bundles" by households. Beginning with data on the standing stock of housing in the previous period, the supply sector models changes in the quality of this stock ("filtering"); alterations of this stock through additions, subdivision, or combination; and the construction of new units on vacant land. The market clearing sector then matches demand for each type of housing bundle with its supply in each residence zone, and adjusts the expected price of this bundle in the direction appropriate for market clearance.

The market clearing mechanism is a linear programming algorithm of the Hitchcock type, which reduces the household location problem to one of cost minimization. To achieve this powerful simplification, some strong assumptions are necessary: (1) Each household has a fixed and known demand for travel to a known set of workplace destinations; this journey-to-work is assumed to dominate the spatial variation in travel cost, which includes a known monetary value-of-time component. (2) Household preferences for a large number of housing attributes, such as number of rooms, size of lot, quality of neighborhood, etc., are treated as preferences for a relatively small number of combinations of these attributes, or "housing bundles." (3) There exist surfaces of market prices for housing bundles which vary by residence location; these prices are assumed known to consumers, who behave as price-takers.

With these assumptions, the NBER model defines for each housing bundle and for each household a surface of gross prices

$$GP_{h,i,j,k,m} = P_{k,i} + (V_h)(H_{i,j,m}) + (C_m)(M_{i,j,m}) \qquad (1)$$

where GP are monthly gross prices, P are monthly bundle prices, V are values of travel time per hour, H are hours of travel per month, C are out of pocket costs per mile, M are the miles of travel per month, h are households, i are residential locations, j are travel destinations (workplaces), m are travel modes, and k are housing bundles.

Gross prices GP are thus the sum of the "contract" or market price P and commuting costs, which include the value of time. The market price P includes a location rent component, an indirect result of the commuting cost and the scarcity of desired housing bundles at convenient locations, but it is only through gross prices that location influences housing demand; there are no specific locational attributes to the desired housing bundles.[3] As the market for each housing bundle is cleared independently of all other housing bundle submarkets, with a separate pass through the linear programming algorithm, the solution yields shadow prices for each bundle type in each residence zone. These shadow prices can then be translated into location rents. The change, if any, in location rent from the previous solution period becomes the change in the expected price of the housing bundle at the residence location, thus moving prices in the appropriate direction for market clearance.[4]

Calibration with Indianapolis Data

The NBER model appears to be under more or less continuous development. Calibration with the latest version, Pittsburgh II, using an especially favorable data set, has encountered many difficulties, although many of these appear to be structural or computational rather than data-dependent. The author's attempt to calibrate the model using Indianapolis data, while premature in some respects, can be justified on several grounds: Despite its imperfect and evolutionary status, the NBER model probably represents the "state of the art" in urban economic models, and should thus see wider testing and application. Feedback from calibration experiments to the model designers should speed convergence to a model design which is sufficiently general.

---

[3] This is equivalent to the assumption that household location involves a two-step decision, in which each household first decides what housing bundle it needs independently of location, and then chooses location solely on cost-minimization criteria.

[4] The objective function currently in use minimizes aggregate transportation costs in each submarket. A possible modification which appears formally equivalent but is substantially closer to actual consumer behavior would be to minimize gross prices.

Model dimensions. The NBER model presents formidable data storage and processing problems. An earlier, more modest version calibrated with Pittsburgh data, known as Pittsburgh I, simulated 96 types of households (family types) employed in 11 industries located at 20 workplaces, and 40 housing bundles available in 50 residence zones. The resulting five-dimensional array contains nearly two million non-empty cells. The current version, Pittsburgh II, includes a further disaggregation of residence zones, which increases the detail to some 7.6 million unique elements. The author's calibration attempt with Indianapolis data was somewhat less ambitious; the data set was constructed for 72 household types employed by 8 industries in 18 work zones, and 12 housing bundles available in 25 residence zones.

Data availability. The choice of the Indianapolis SMSA for the calibration experiment was based on three criteria: (1) Proximity of the test city to the experimenters; (2) Existence of a recent transportation study of the test city; and (3) Absence of any extreme characteristics in the city's socio-economic data set. An indispensable data input to any urban model in which the journey-to-work matters are data on the workplace locations of households, by residence zone. If, as in our case, an origin-destination study is available, one can make at least a first approximation to the work trip data. There are no Census data of this type, and the scope and methodology of transportation studies are highly variable across cities. The Indianapolis transportation study, for example, generated interzonal trip data, but did not determine the purpose of each trip. This obstacle to calibration is characteristic of much of the required data, which, if available at all, comes from a study unique to the city in question. To create the Indianapolis work trip data, we simply assumed that all journeys to the 18 designated work zones were work trips. Since the transportation study defines its own set of geographic regions, many of which are not congruent with Census tracts, meshing of the two data sets not only involved some crude approximations but also led us to draw workplace and residence zone boundaries so as to maximize congruence, another instance of the data set influencing model structure.

Other holes in the input data set were filled by assuming that an unknown distribution of an aggregate across subsets would be identical to a known distribution for some other city. For example, the number of workers at each workplace was known, but its stratification across the 72 classes of households was not available. To create the data set, we had to assume that the proportion of households in each household class was the same as in Detroit, for which a detailed study had been made. This is equivalent to assuming that Detroit and Indianapolis have the same or very similar socio-economic structures, a hypothesis which, like so many others in the experiment, was probably untenable but had to remain untested. Similarly, mobility rates by household class were taken to be identical to those measured in a study of the San Francisco Bay Area.

Probably the most serious data deficiency was the absence of a sufficiently detailed market price surface, $P_{k,i}$ in Equation (1). The entire theoretical structure of the model rests upon the "gross price hypothesis" implicit in this equation. An early step in calibration should thus be the statistical testing of this hypothesis through the estimation of relations, in which demand for housing bundles of each type is a function of gross prices, household income, and other household characteristics. The resulting coefficient matrix is essential to the simulation. Instead, we took the existing matrix of demand coefficients from Pittsburgh II and relied upon the model itself to generate the price detail lacking in the Census data! While this procedure may yet produce some interesting results, it can hardly be recommended as a calibration method.

There were a few bright spots. The "Filtering" submodel of the NBER model, which determines the rates at which existing structures are converted to different kinds of housing, requires data on conversion costs. This transformation matrix was constructed with some confidence from cost data published by the construction industry for a large set of cities, including Indianapolis.

Perhaps the most sensible reaction to this accumulation of data deficiencies (of which I have described only a few) in the Indianapolis data set would be to abandon the calibration experiment. This is essentially what we were forced to do, although the model itself remains under study. An alternative would be to alter the structure of the model in order to conform to the available data. In the absence of an econometric study of housing market behavior in Indianapolis, of the kind suggested earlier, a drastic simplification of model structure would be required. A third possibility would be to generate the necessary data by conducting the appropriate set of supporting studies; should resources for this purpose become available, this would be attractive, but it begins to stretch the meaning of calibration.

IV. CONCLUSION

Structural limitations

As with many other large scale simulation models, the designers of the NBER model look to further disaggregation of data and more model detail as the main road to improvement in model performance. In particular, the urban housing stock is seen as being more heterogeneous than the current version represents. Why not increase the model dimensions by further disaggregation of residence zones, housing bundles, household types, etc.? In Pittsburgh II, this process seems to be approaching a kind of upper bound on structural detail, imposed by the Census disclosure rules.[5] If we may venture a supposition and a loose analogy, this poorly-defined upper bound may impose a fundamental limit on the measurement accuracy of social science models, suggestive of the Heisenberg Uncertainty Relation. Were society willing to revoke the disclosure rules, the limitation would not necessarily disappear; in the "glare of publicity," individual behavior might well be affected by the model builder's measurement effort.

Some Design Guidelines

The essentially unsuccessful attempt reported here to calibrate the NBER Urban Simulation Model with Indianapolis data may at least yield a few rather obvious design guidelines, for the possible benefit of future efforts to construct general urban models.

1. Apply the test of wide data availability and consistency of definition to all variables used in the model. In many cases, this will mean increased reliance on major federal sources such as the Census. If necessary, alter the model structure so as to conform with the available data.
2. During the construction of the model, use analysis of variance or other appropriate econometric techniques to test the validity of

---

[5] The Census and other data-publishing government agencies are required by law not to disclose information about specific individuals, households, or firms. The practical result is that cells of any data stratification which represent aggregates of three or fewer members will not be published or made available to researchers.

structural hypotheses over a wide cross section of sampled cities, rather than just one or two.

3. Be careful that the urban model is not so disaggregated as to run seriously afoul of the disclosure rules, especially when applied to small or unusual cities.

REFERENCES

[1] Goldstein, Gerald S., and Leon N. Moses. "A Survey of Urban Economics," Journal of Economic Literature, 11, 2 (June 1973), 471-515.

[2] Ingram, Gregory K., John F. Kain, and J. Royce Ginn. The Detroit Prototype of the NBER Urban Simulation Model (New York: National Bureau of Economic Research and Columbia University Press, 1972).

[3] Judge, George G. On Estimating the Parameters of Economic Relations: A Review (Urbana: College of Commerce and Business Administration - University of Illinois at Urbana-Champaign, 1973).

[4] Kain, John F., and John R. Meyer. "Computer Simulations, Physio-Economic Systems, and Intra-Regional Models," American Economic Review, 58, 2 (May 1968), 174-181.

[5] Mills, Edwin S. Studies in the Structure of the Urban Economy (Baltimore: Johns Hopkins Press, 1972).

[6] Nagelberg, Mark. "Simulation of Urban Systems - A Selected Bibliography" IFF Working Paper WP-3 (Middletown, Connecticut: Institute for the Future, Inc., 1970).

SOCIO-ECONOMIC INFLUENCES AND INTERACTIONS WITH TRANSPORTATION SYSTEMS

HRAIR ALDERMESHIAN and MARLIN H. MICKLE
University of Pittsburgh
Pittsburgh, Pennsylvania

ABSTRACT

The systems approach to socio-economic modeling has gained considerable attention in recent years. Industry, society, housing, pollution, capital, etc. have been included in these modeling efforts. An extension to a transportation model is presented in this paper, which takes into consideration interactions and influences of pollution, social attitudes and personal psychology.

BACKGROUND

The socio-economic systems area has received a considerable amount of attention in recent years. Basically, there are two popular methods of attack in the modeling of socio-economic systems. One method is essentially characterized by the works of Forrester and Meadows[1,2,3,4,5]. This method is one of proceeding from the general to the specific with a particular concern for general trends and is not primarily concerned with numerical verification. The second method is essentially characterized by the works of Vogt and Mickle[6,7,8,9,10,11]. This method is one of identifying those aspects of the disciplines which are a part of the socio-economic modeling effort which can be mathematically modeled and for which at least a certain amount of verification is possible both structurally and numerically. Both methods have their utility, and neither can be recommended to the exclusion of the other.

Later efforts of both Forrester and Meadows[3,4,5] were extended to include aspects of pollution, capital, etc. in the development of the world models. One of the aspects of our society which is of current concern is transportation. While the lack of certain types of transportation is of particular concern, especially to those who need it, transportation is also an important facet of our society which encompasses a number of aspects including capital, pollution, energy, social stratification, etc. with a number of influences and interactions.

A socio-economic model can not completely describe a society without incorporating at least some aspects of transportation. While much work has been published on many facets of transportation, the two major modeling efforts mentioned previously have not considered transportation in this particular light. Because of the validity of both approaches to socio-economic modeling, certain aspects of transportation have been included in one model presented in this paper and illustrate the application of the Forrester and Meadows approach to modeling socio-economic systems.

The aspects of capital, pollution, social attitudes and personal psychology are of importance to our society and are an integral part of the area of transportation. While many efforts have been put forth to incorporate and identify these aspects and their interactions with transportation, the efforts have not been concerned with specific tractable models which exhibit the above aspects and their interactions and interrelations. These aspects are included in the model presented in this paper.

The methodology of the type used by Forrester and Meadows is a valuable tool for modeling certain systems. Thus, the present approach to transportation modeling is that of systems dynamics with a corresponding DYNAMO simulation program. The second approach, the dynamic system modeling method of Vogt and Mickle, is used to model socio-economic influences and interactions with transportation in a separate work[12] which will be published at a later date. While both models have similar aspects and variables of concern, no effort is made to equate the models themselves or evaluate any numerical results. Instead, such parallel development is intended to be illustrative of the methods of approach to the modeling of aspects of socio-economic systems.

INTRODUCTION

In the past decade, socio-economic effects of transportation pollutants have rapidly advanced to the foreground of public attention, as the property and health hazards they cause have become more tangible. When President Johnson signed the Clean Air Act in 1963, he expressed the nation's concern when he stated, "We have now reached the point where our factories, our automobiles, our furnaces, and our municipal dumps are spewing out more than 150 million tons of pollutants into the air we breath - almost one-half million tons a day."

Considering New York City as an example, the magnitude of pollutants due to automotive transportation amounts to 695 tons per day[13]. This enormous pollution problem is pitted against the need for better transportation for an ever increasing number of commuters from the suburbs to the core of the city.

Medical researchers have established a definite pattern between increases of the cardiorespiratory diseases and air pollution[14,15]. Despite the critical nature of the problem, however, the haphazard control measures presently in effect emphasize the need for a comprehensive study of various policies of pollution control for automotive transportation.

Is it, for instance, effective to control automobile exhausts at a certain level without regard to exponential growth of commuters to the inner city? How does an increase in parking rates affect the modal split of transportation between private automobile and public bus transportation?

To investigate these questions, a model containing interactions of transportation, pollution and the effects of pollution on human beings was developed and a subsequent study was undertaken. A description of the model and some of the results of the study are included in this paper.

THE MODEL

The model assumes a single modal split of commuters between automobile and bus transportation. The transportation network leading to the city core is assumed to be of zero width and aggregated to represent total accessibility to the inner city. While for any particular city, such an assumption may not be valid, the particular configuration is city or area dependent. Such extensions are however straightforward and follow from the example presented here which is illustrative without undue complication. In the model, commuters are assumed to make single trips to and from the city daily.

The modal split is effected by considering the cost of a number of facets of a given trip such as the period of wait for public transportation, the travel cost, the parking cost for private transportation, the cost of time lost in transit which are a part of the expense for the two modes of transportation. The nonlinear interrelationships assumed for the indicated model variables of the model are shown in Figure 1.

In the model, the preference of people for commuting by car rather than by bus is implicitly included in the calculations of the change of the rate of bus commuters by assuming that if the cost of travel by car is equal to the cost of travel by bus, people prefer to use cars. Thus, in this case, the change of the rate of bus commuters is zero.

The rate of increase of commuters over the model run is assumed to be exponential and is provided in table form. The adverse effect of pollution on the commuters' health condition is also included. The alarming increase of death-rates for degenerative heart diseases, lung cancer and, bronchitis and emphysema is represented in Figure 2[14]. While the total increase may be due to causes other than pollution, a definite trend is discernable. In 1965, for example, 3.2 million in-hospital days were recorded in Canada for respiratory diseases[14]. The indicated situation is not unlike that of the world models in The Limits to Growth[4] and The Dynamics of Growth in a Finite World[5].

The feedback-loop structure of the model is shown in Figure 3. A positive feedback-loop of exponential growth is represented for the increased rate of commuters by the commuter increase rate, COMMR. A feedback-loop of a goal seeking nature is the effect of pollution on the health of commuters. An increase in commuters will result in an increase of pollution when no effective means are employed to decrease this trend. The increase in pollution will lead to a decrease of commuters due to deaths or hospital confinements. A negative feedback-loop of pollution control is incorporated in the model. An increase in technological control of pollution will increase the rate of removal of pollutants, POLRMV, and thus effect a desired pollution level. Another negative feedback-loop incorporates the level of commuters by bus. In this case, an increase in bus commuters will lead to a decrease of density of traffic and consequently an improvement of travel conditions and pollution. Thus, commuters using private automobiles will increase reversing the above trend. Additional details of the model can be observed in the DYNAMO[16] flow chart of Figure 4.

MODEL RUNS

In the model runs, the plots identify the variables by a single letter; for example, pollution, POLLU, is identified by P. The vertical axis is common to all variables while the horizontal axis varies from one variable to another.

In Run 1, no pollution control policies are implemented, pollution removal is effected only by natural means. Pollution grows exponentially leading to a disasterous effect on the commuting population. After this disaster, there is a shift from automobile commuting to bus commuting but the pollution effects predominate.

In Run 2, a given standard of pollution control is introduced in the year 2000. Effects similar to the first run result. They are delayed however and are somewhat less awesome. As pollution increases, the step input of control regulates only the more discernable of pollutants while the fine particulates and scarce pollutants are left uncontrolled leading to the model behavior.

In Run 3, the parking fee for cars is increased drastically. The car commuter rate is decreased slightly, but the pollution is not appreciably changed, and the disaster of Run 1 is prominent after the year 2300.

CONCLUSIONS

The purpose of developing this particular model has been to examine the inherent trends evidenced by the influences and interactions involving transportation pollutants and certain socio-economic factors.

The model should in no way be used to predict exact dates and numbers concerning future events.

The results obtained using the present model definitely underline the need for determined steps to curb pollution now, before the problem becomes uncontrollable.

The methodology of dealing with socio-economic systems used in this paper, frequently termed systems dynamics, is effective in socio-economic modeling and represents one method of modeling such systems.

REFERENCES

1. Forrester, J., Industrial Dynamics, MIT Press, 1961.
2. Forrester, J., Urban Dynamics, MIT Press, 1969.
3. Forrester, J., World Dynamics, Wright-Allen Press, 1971.
4. Meadows, D., et al, The Limits to Growth, Universe Books, 1972.
5. Meadows, D., et al, The Dynamics of Growth in a Finite World, to be published.
6. Vogt, W.G., M.H. Mickle and H. Aldermeshian, "A Dynamic Input-Output Model for an Economic System," Proceedings of the Ninth Allerton Conference on Circuit and System Theory, 1971.
7. Vogt, W.G., and M.H. Mickle, "Labor in a Dynamic Economic Model," Proceedings of the Fifth Annual Princeton Conference on Information Sciences and Systems, 1972.
8. Mickle, M.H. and W.G. Vogt, "A Dynamic Labor Allocation Model Including Efficiency," Proceedings of the Tenth Allerton Conference on Circuit and System Theory, 1972.
9. Vogt, W.G. and M.H. Mickle, "Stochastic Model for a Dynamic Leontief System," Proceedings of the Tenth Allerton Conference on Circuit and System Theory, 1972.

10. Mickle, M.H. and W.G. Vogt, "A Dynamic Sociological Model," Proceedings of the IFAC-IFORS Conference on Systems Approaches to Developing Countries, 1973.

11. Vogt, W.G., M.H. Mickle, H. Aldermeshian, K. Amirebrahimy, A. Nauda and T.A. Trygar, "A Dynamic Housing Simulation Model," to be published as a paper in the NATO ASI Book Series, 1973.

12. Mickle, M.H., W.G. Vogt, H. Aldermeshian and T.A. Trygar, "An Input-Output Model of Socio-Economic Influences on Transportation," to be published.

13. Council of the City of New York, "Air Pollution in New York," Report M-970, 1965.

14. DeVilliers, A.J., "Effects of Air Pollution on Health," Man and His Environment, Vol. 1, Pergamon Press, 1970.

15. Ayres, S.M., "The Effects of Air Pollution Upon Health," Proceedings of the Rutgers Seminar, 1967.

16. Pugh, A.L., DYNAMO II, User's Manual, 1973.

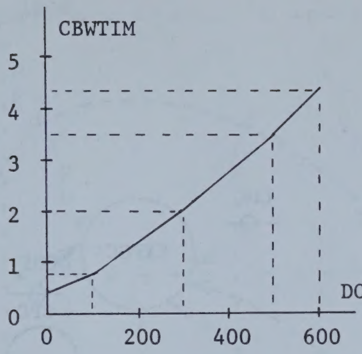


Figure 1a. CBWTIM vs. DOTF

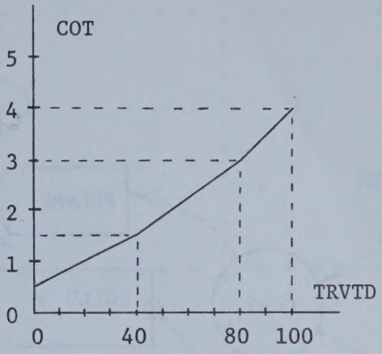


Figure 1b. COT vs. TRVTD

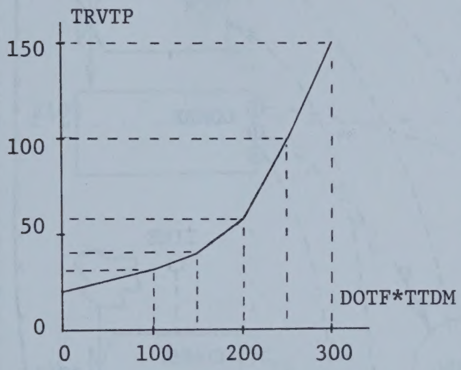


Figure 1c. TRVTP vs. DOTF*TTDM

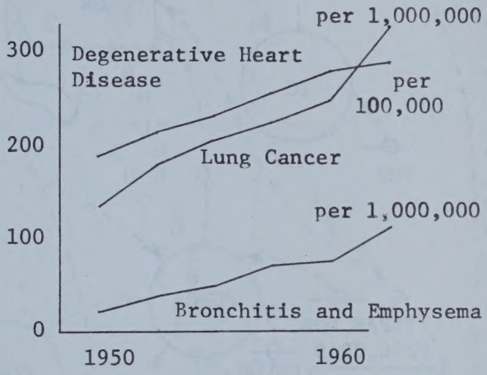


Figure 2. Age Standardized Death Rate (Males)

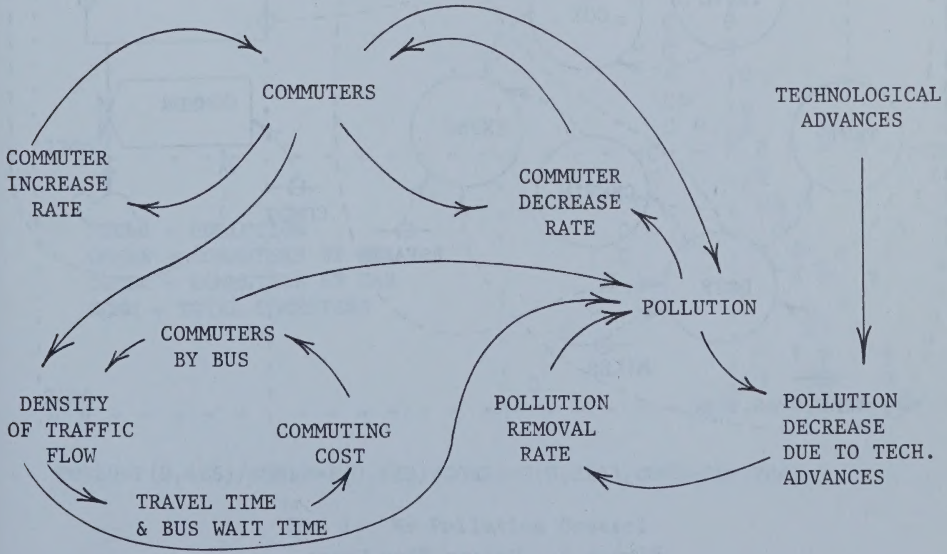

COMMUTERS
TECHNOLOGICAL ADVANCES
COMMUTER INCREASE RATE
COMMUTER DECREASE RATE
POLLUTION
COMMUTERS BY BUS
DENSITY OF TRAFFIC FLOW
COMMUTING COST
POLLUTION REMOVAL RATE
POLLUTION DECREASE DUE TO TECH. ADVANCES
TRAVEL TIME & BUS WAIT TIME

Figure 3. Feedback - Loop Structure

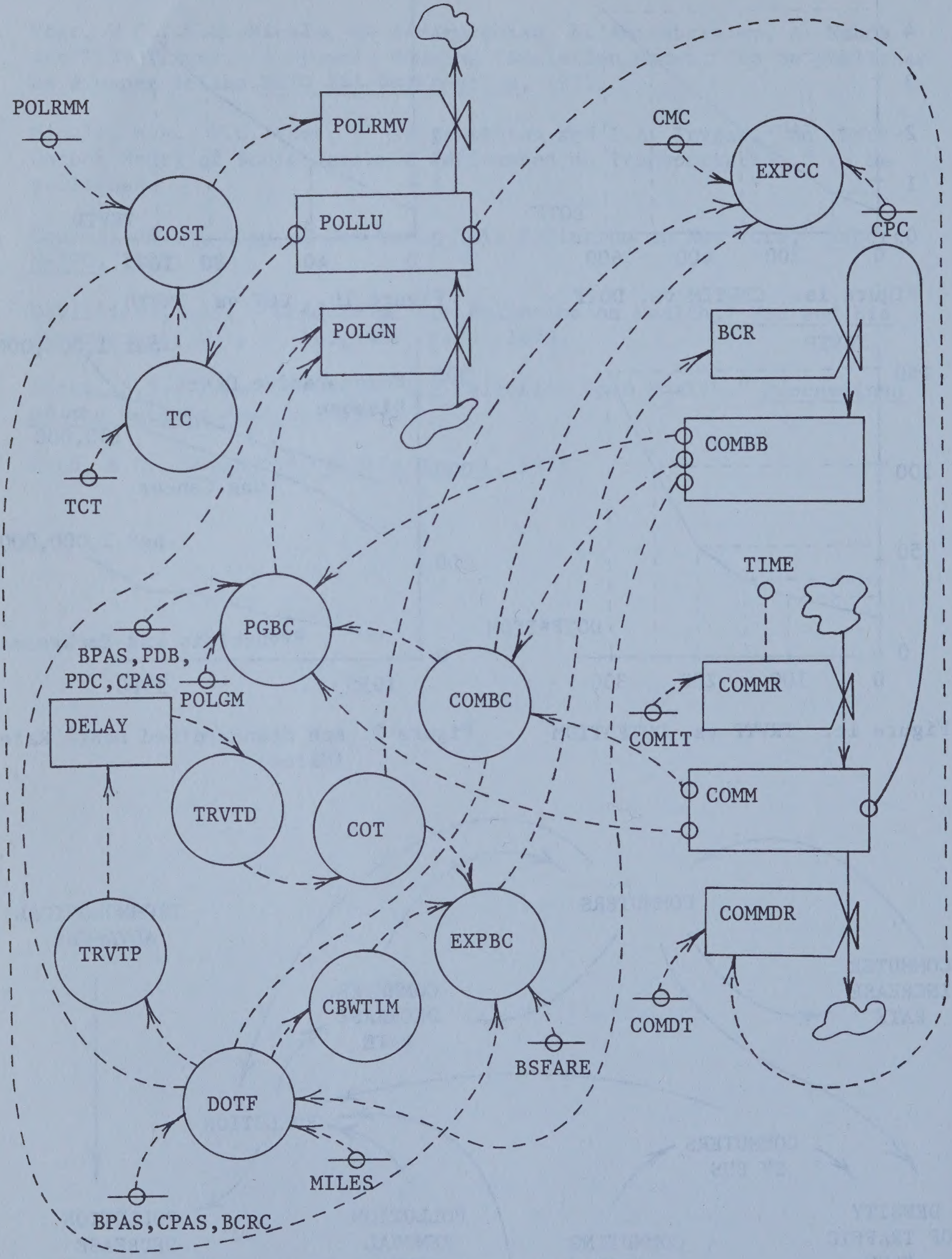
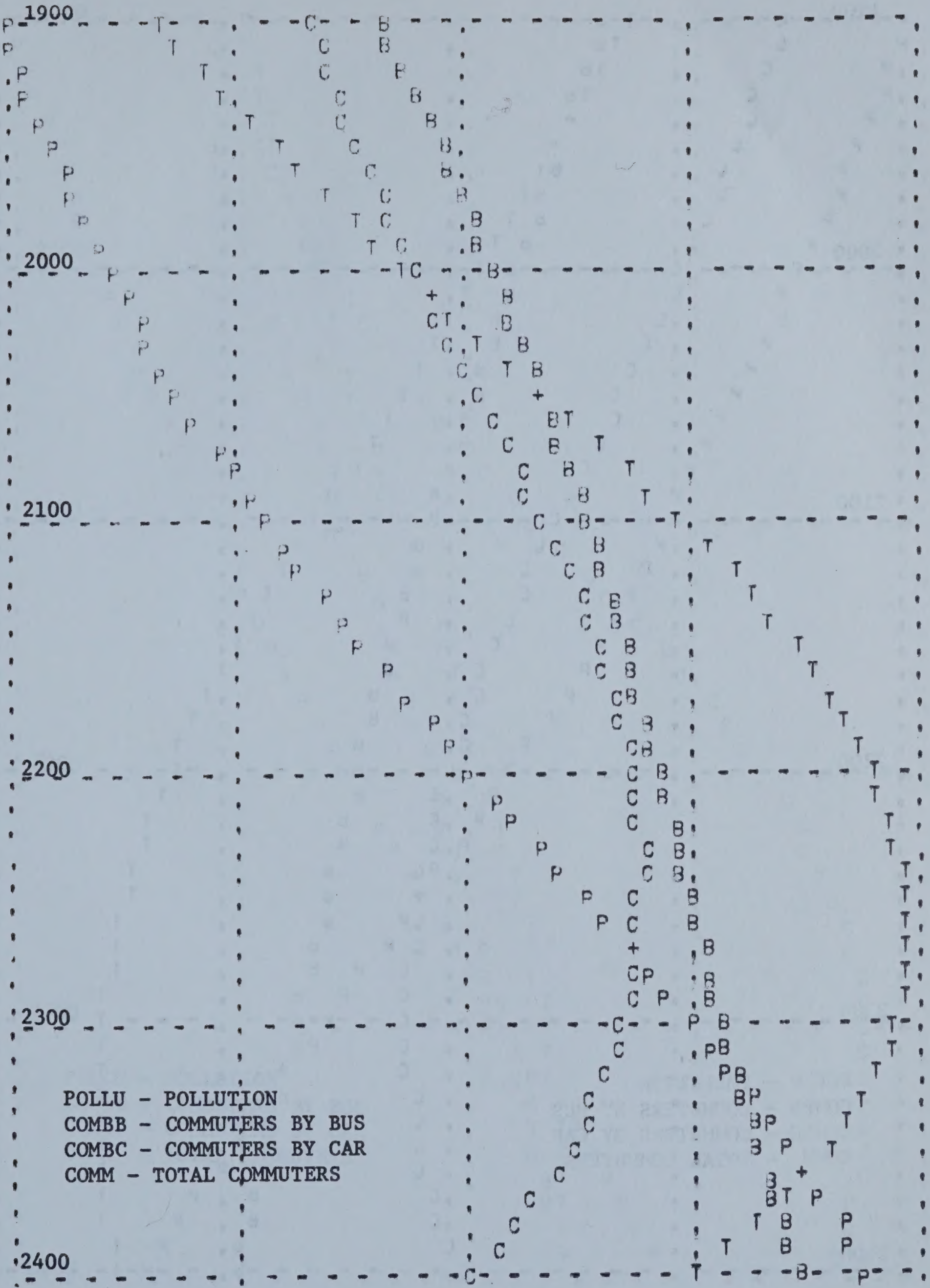
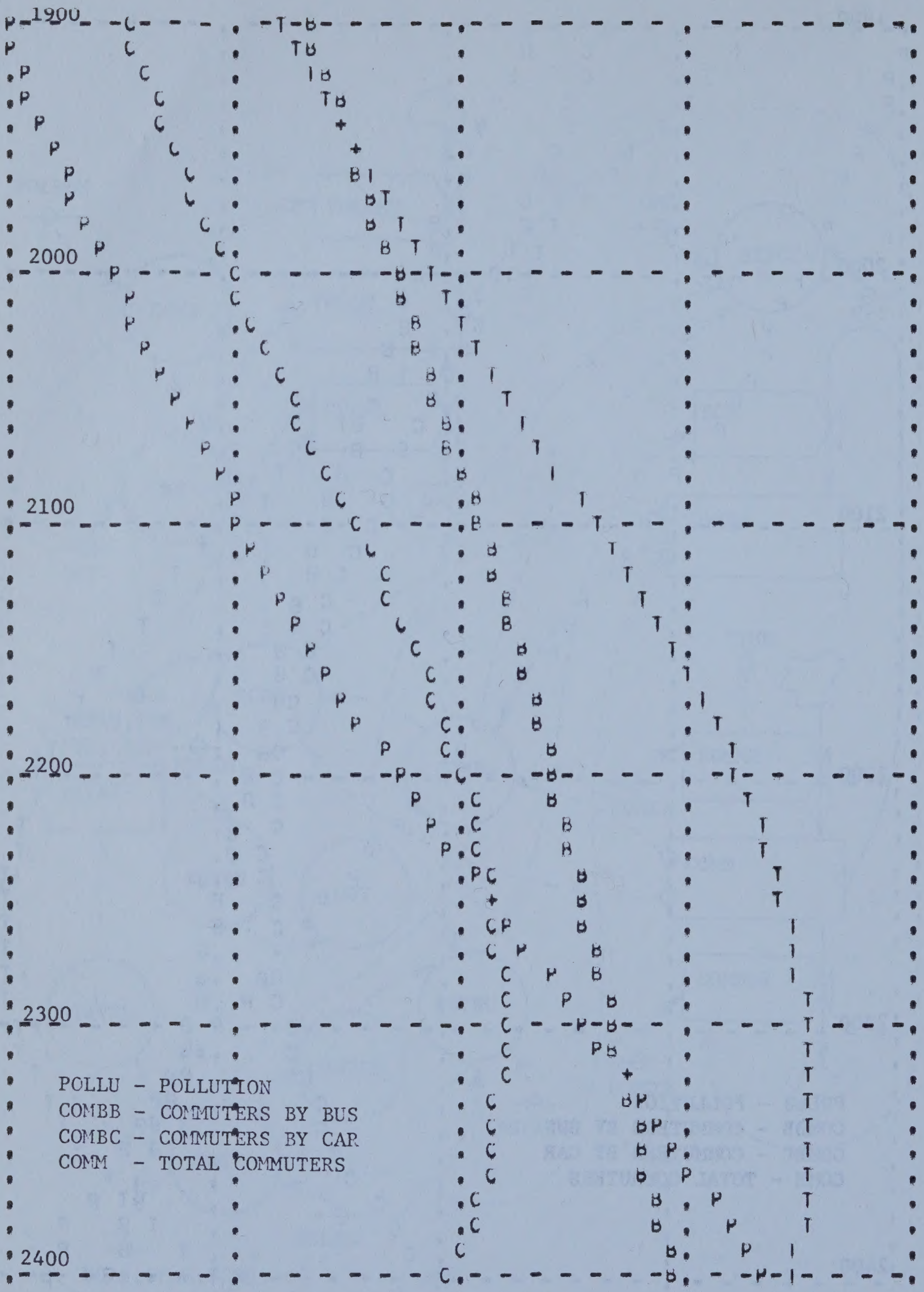

POLRMM
POLRMV
CMC
EXPCC
POLLU
COST
CPC
POLGN
TC
BCR
COMBB
TCT
TIME
PGBC
BPAS,PDB,
PDC,CPAS
POLGM
COMBC
COMMR
DELAY
COMIT
TRVTD
COT
COMM
EXPBC
COMMDR
TRVTP
CBWTIM
COMDT
BSFARE
DOTF
MILES
BPAS,CPAS,BCRC

Figure 4. Dynamo Flow Chart



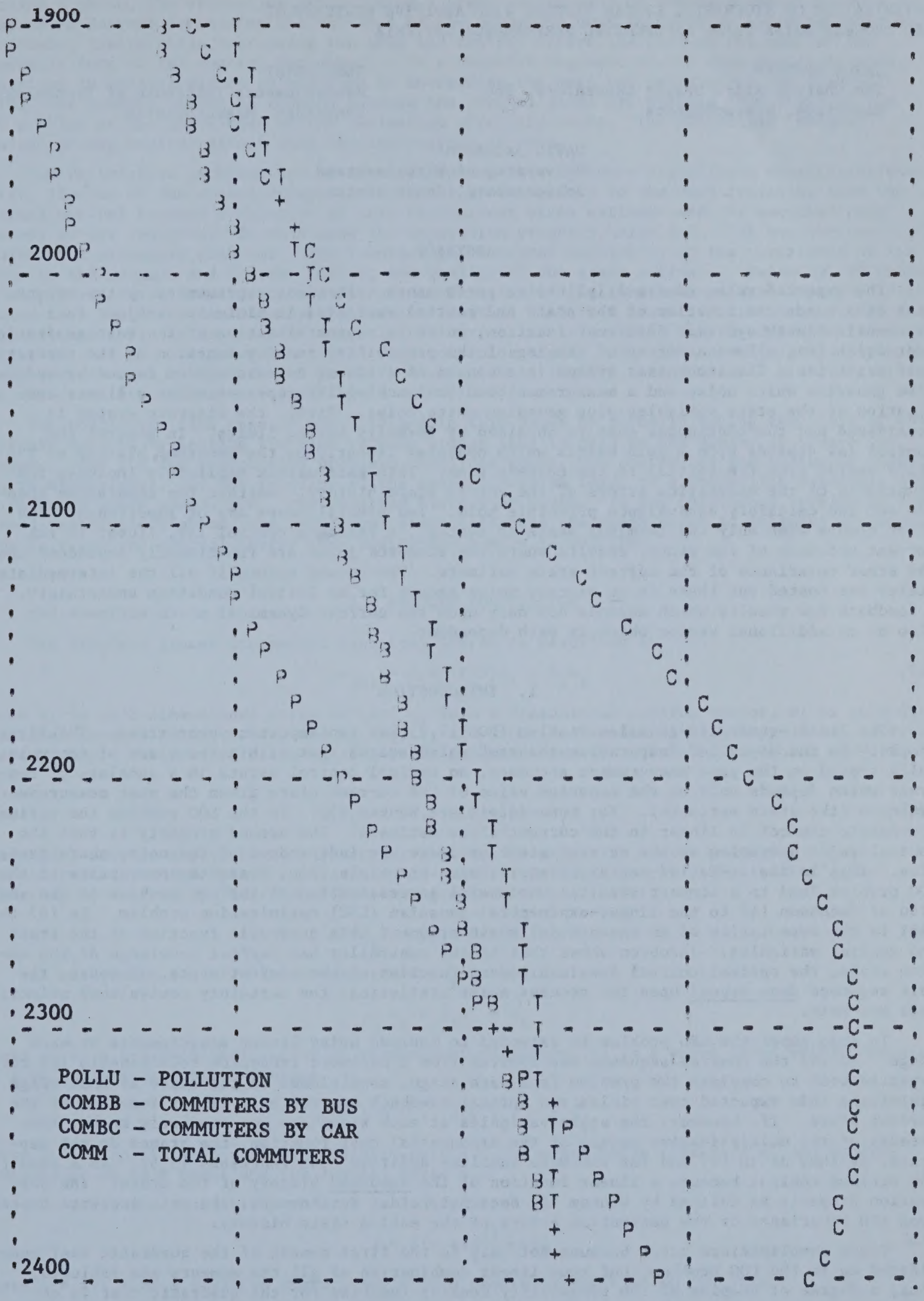
POLLU=P(0,4E6)/COMBB=B(0,8E3)/COMBC=C(0,2E4)/COMM=T(0,2E4)

Run 1: No Pollution Control



POLLU=P(0,4E6)/COMBB=B(0,1E4)/COMBC=C(4E3,2.4E4)/COMM=T(4E3,2.4E4)

Run 2: Constant Pollution Control



POLLU=P(0,4E6)/COMBB=B(0,2E4)/COMBC=C(5E3,1.3E4)/COMM=T(0,4E4)

Run 3: Increase in Parking Fare

OPTIMIZATION OF STOCHASTIC LINEAR SYSTEMS WITH ADDITIVE MEASUREMENT AND PROCESS NOISE USING EXPONENTIAL PERFORMANCE CRITERIA

JASON SPEYER*
The Charles Stark Draper Laboratory, Inc.
Cambridge, Massachusetts

JOHN DEYST
Massachusetts Institute of Technology
Cambridge, Massachusetts

DAVID JACOBSON†
University of Witwatersrand
Johannesburg, South Africa

ABSTRACT

The expected value of a multiplicative performance criterion, represented by the exponential of a quadratic function of the state and control variables is minimized subject to a stochastic linear system. This cost function, which is a generalization of the mean quadratic cost criterion, allows a degree of shaping of the probability density function of the quadratic cost criterion. The stochastic system is composed of a linear dynamic system forced by additive gaussian white noise and a measurement vector functionally represented as a linear combination of the state variables plus gaussian white noise. First, the discrete system is considered and the continuous case is obtained by formally taking limits. In general the control law depends upon a gain matrix which operates linearly on the smoothed history of the state vector from the initial to the current time. This gain matrix explicitly includes the covariance of the estimation errors of the entire state history. Neither the separation theorem nor the certainty equivalence principle hold. Two special cases are of importance. The first occurs when only the terminal state is costed. A feedback control law, linear in the current estimate of the state, results where the feedback gains are functionally dependent upon the error covariance of the current state estimate. The second occurs if all the intermediate states are costed but there is no process noise except for an initial condition uncertainty. A feedback law results which depends not only upon the current dynamical state estimate but also on an additional vector which is path dependent.

1. INTRODUCTION

The linear-quadratic-gaussian problem (LQG){1,2} has two important properties. The first property is the so-called "separation theorem" which states that within the class of controls which depend on the past measurement sequence, an optimal control exists in a subclass of controls which depends only on the expected value of the current state given the past measurement sequence (the state estimate). For more detail see Wonham {3}. In the LQG problem the optimal stochastic control is linear in the current state estimate. The second property is that the control gains operating on the current state estimate are independent of the noise characteristics. This is the so-called certainty equivalence principle {2}. These two properties of the LQG problem lead to a compact result. One useful generalization of the LQG problem is the solution of Jacobson {4} to the linear-exponential-gaussian (LEG) optimization problem. In {4} the cost is the expectation of an exponential whose argument is a quadratic function of the state and control variables. Jacobson shows that if the controller has perfect knowledge of the entire state, the optimal control law is a linear function of the current state. However, the gain sequence does depend upon the process noise statistics; the certainty equivalence principle does not hold.

In this paper the LEG problem is extended to include noisy linear measurements at each stage. In {4} the control sequence was derived from a backward recursion relationship for the expected cost to complete the problem from each stage, conditioned on the state at that stage. Minimizing this expected cost yields the optimal feedback control as a linear function of the current state. If, however, the state variables at each stage are not precisely known, then because of the multiplicative nature of the exponential cost function, the stages do not separate, as they do in {4} and for the more familiar additive cost functions {3,5}. As a result, the optimal control becomes a linear function of the smoothed history of the state. The separation property as defined by Wonham {3} does not hold. Furthermore, the gain sequence depends upon the covariance of the estimation errors of the entire state history.

These complications occur because not only is the first moment of the quadratic cost considered as in the LQG problem, but some linear combination of all the moments are included. Thus, a degree of shaping of the probability density function for the quadratic cost is obtained. The exponential cost function is viewed as a generalization of the quadratic cost function for which the LQG problem results as a limiting case. As the one stage example of

---

*On leave 1972-73, Dept. of Applied Mathematics, The Weizmann Institute of Science, Rehovot, Israel.

†This author's research was partially supported by the joint services electronics program at Harvard University on Contract No. 0014-67-A-0298-006.

Section 4 shows, the controller design using a positive argument of the exponential becomes a compromise between decreasing the variance of the quadratic cost in order to establish narrower confidence limits while increasing the mean and control effort (defined as the mean of the quadratic form of the control variables). If a negative argument of the exponential is used, reduction in control effort is achieved by increasing the mean and variance of the quadratic form. The control effort is reduced because the control gains are reduced from those of the LQG problem as the covariance of the estimation error increases. The controller "hedges" against strong control effort when the information is poor.

Two restrictive problems are of interest because they produce significant simplifications. First, if none of the states except the terminal state appears in the cost function, then the optimal control becomes a function of only the current state estimate and the smoothed past history is not required. In this case the separation property holds (cf. {3}) but the certainty equivalence principle does not. The feedback gains depend explicitly on the covariance of the error in the estimate and thereby reflect the quality of the state estimate. Secondly, if there is no process noise driving the system, a cost function of the intermediate states can be transformed to a cost function of only the terminal state. With this transformation the optimal control becomes a linear function of not only the current state estimate but an additional state vector defined as a weighted sum of the past control actions. This is a particularly simple example where both the separation property does not hold (due to the additional state vector) and the certainty equivalence principle does not hold.

The continuous time results are obtained in this paper by formally taking the limit of the discrete results by letting the time between stages tend to zero as the number of stages increases. For the two restrictive problems, outlined above, the controller and estimator remain finite dimensional, but the solution to the general problem leads to a controller dependent upon an infinite state space, i.e., the smoothed history of state. In the continuous version of the second restricted problem, the additional state variables are reminiscent of integral compensation in classical control theory when additional memory is required.

2. PROBLEM FORMULATION

The discrete linear stochastic dynamical system is described by

$$x_{i+1} = A_i x_i + B_i u_i + \Gamma_i w_i \tag{1}$$

where x_i is an n-dimensional state vector, u_i is a p-dimensional control vector, w_i is an m-dimensional process noise vector, and A_i, B_i, Γ_i are known, nonrandom (n×n), (n×p), (n×m) matrices, respectively. The process noise w_i is a white gaussian sequence of random variables with statistics

$$E(w_i) = 0 \; , \quad E(w_i w_j^T) = w_i \delta_{ij} \tag{2}$$

where E( ) denotes expectation, superscript T indicates matrix or vector transpose, δ_{ij} is the Kronecker function and $W_i \geq 0$ indicates a non-negative definite matrix. The initial state vector x_1 is a normally distributed random variable with statistics

$$E(x_1) = 0 \; , \quad E(x_1 x_1^T) = Y_1 \tag{3}$$

At each stage a noisy linear measurement z_i is available to the controller

$$z_i = H_i x_i + v_i \tag{4}$$

where z_i is a q-dimensional vector of measurement variables, H_i is a known (q×n) matrix and v_i is a white gaussian noise sequence

$$E(v_i) = 0 \; , \quad E(v_i v_j^T) = v_i \delta_{ij} \tag{5}$$

where $V_i > 0$ indicates a positive definite matrix. The white noise processes w_i and v_i are independent of each other and the initial state x_1.

The performance criterion or cost to be minimized is

$$J = E(\mu \exp\{\tfrac{1}{2} \mu\psi\}) \tag{6}$$

where $\mu = \pm\gamma$; γ is a positive scalar; and ψ is a quadratic function of the form

$$\psi = \sum_{i=1}^{N} (x_i^T Q_i x_i + u_i^T R_i u_i) + x_{N+1}^T Q_{N+1} x_{N+1} \tag{7}$$

where N+1 is the final stage, $R_i \equiv R_i^T > 0$ and $Q_i \equiv Q_i^T \geq 0$. Note that the cost is multiplicative since the sum in (7) appears as the argument of the exponential in (6). This class of cost functions was explicitly excluded in the work of {3,5}.

The choice of μ in (6) determines the type of cost function to be minimized. If $\mu = +\gamma$, the cost on large values of the state and control is very heavy. Expanding the exponential in (6) in a Taylor series yields

$$J = \gamma(1 + \gamma E(\psi)/2 + \gamma^2 E(\psi^2)/4 \cdot 2! + \gamma^3 E(\psi^3)/8 \cdot 3! + \ldots) \tag{8}$$

The cost is a particular weighted sum of all the moments of ψ with γ as the only free parameter. This allows a degree of shaping of the probability density function for the quadratic cost. If all the moments of ψ remain bounded as γ becomes arbitrarily close to zero, the first moment dominates the sequence and the optimization problem specializes to the standard LQG problem. If $\mu = -\gamma$, the cost resembles a likelihood function where large excursions in the state and control variable are not costed very heavily. The cost function of (6) using $\mu = -\gamma$ might represent, approximately, the probability that the entire state and control history lie within specified bounds, and the solution may be useful for designing systems to operate within specified bounds or limits.

The stochastic optimization problem is to minimize the expected cost J given by (6) and (7) subject to the stochastic system (1) and (4) using a nonanticipative control vector which may be a function of the entire past measurement history. The admissible controls are required to be Borel functions of the past measurements to assure that they are random variables. Define the measurement history as a composite vector

$$Z_i^T = (z_1^T, z_2^T, \ldots, z_i^T) \tag{9}$$

which is an element of $R^{i \cdot q}$. The control u_i may be any Borel function of Z_i denoted as $u_i(Z_i)$.

Similarly, define Ξ_i as the smallest σ algebra with respect to which Z_i is measurable. Furthermore, define $E_{Z_i}(\cdot)$ as the expectation conditioned on the σ algebra Ξ_i. The conditional expectation $E_{Z_i}(\cdot)$ is itself a random variable which, by definition, is measurable with respect to Ξ_i and is a Borel function of Z_i (cf. Wong {6}).

The cost function (6) can be written in terms of nested conditional expectations. Since Ξ_i is the smallest σ algebra with respect to which Z_i is measurable, the σ algebra Ξ_{i+1} contains the algebra Ξ_i, i.e., $\Xi_{i+1} \supset \Xi_i$, and the cost may be written as a sequence of nested conditional expectations

$$J = E(E_{Z_1}(E_{Z_2}(\ldots E_{Z_{N+1}}(\mu e^{\mu\psi/2}) \ldots))) \tag{10}$$

If the cost to go to the final stage from stage i, given Z_i, is defined as

$$J_i(Z_i) = E_{Z_i}(\mu e^{\mu\psi/2}), \tag{11}$$

where it is assumed that an admissible control sequence has occurred up to stage i-1, then a recursion formula for $J_i(Z_i)$ for all $i \epsilon (1,N)$ is obtained directly from (10) and (11) as

$$J_i(Z_i) = E_{Z_i}(J_{i+1}(Z_{i+1})) \tag{12}$$

This recursion formula plays a large part in the derivation of the optimal stochastic controller.

3. SOLUTION OF THE STOCHASTIC OPTIMAL CONTROL PROBLEM

In solving the general exponential cost problem, the optimal stochastic controller is easily obtained from the recursion formula (12) if the problem is reformulated as a terminal cost problem ($Q_i \equiv 0$ $i \epsilon [1,N]$). Therefore, the terminal cost problem is solved first. The general problem is then cast into a terminal cost problem by defining a state as the composite vector of the state history up to the current stage. Finally, it is shown that in the case of only measurement noise but no process noise, the state vector of the general solution collapses to a 2n dimensional state vector.

3.1 Terminal Cost Problem

The terminal cost problem is defined by assuming that in the quadratic argument (7) at least one of the eigenvalues of Q_{N+1} is nonzero and $Q_i \equiv 0$ for $i \epsilon [1,N]$. The terminal condition on the recursion formula (12) using (11) becomes

$$J_{N+1}(Z_{N+1}) = E_{Z_{N+1}}\left[\mu \exp\left\{\frac{\mu}{2}\left(\sum_{i=1}^{N} u_i^T R_i u_i + x_{N+1}^T Q_{N+1} x_{N+1}\right)\right\}\right] \tag{13}$$

Since u_i is by definition, a Borel function of Z_i, it is measurable with respect to Ξ_i. Hence, $u_i(Z_i)$ is measurable with respect to Ξ_{N+1} because $\Xi_{N+1} \supset \Xi_i$. Thus

$$J_{N+1}(Z_{N+1}) = \mu \exp\left\{\frac{\mu}{2} \sum_{i=1}^{N} u_i^T R_i u_i\right\} E_{Z_{N+1}}\left[\exp\left\{\frac{\mu}{2} x_{N+1}^T Q_{N+1} x_{N+1}\right\}\right] \tag{14}$$

Since the expectation in (14) is over a function of x_{N+1} given Z_{N+1} the probability density function for x_{N+1} given Z_{N+1}, $p(x_{N+1}/Z_{N+1})$, is desired. It is well known {1,2} that for the stochastic system of (1) and (4), at any stage i, $p(x_i/Z_i)$ is normal with conditional mean $\hat{x}_i \equiv E_{Z_i}(x_i)$, propagated by the linear difference equation as given by Kalman {7} which is the weighted combination of the extrapolated estimate of x_i, $\bar{x}_i$, and the new information as

$$\hat{x}_i = \bar{x}_i + K_i s_i \tag{15}$$

where $\bar{x}_i$ and the measurement residual, s_i, are

$$\bar{x}_i = E_{Z_{i-1}}(x_i) = A_{i-1}\,\hat{x}_{i-1} + B_{i-1}\,u_{i-1}\,, \quad s_i = z_i - H_i\bar{x}_i \tag{16}$$

The weighting K_i is the Kalman filter gain given as

$$K_i = P_i H_i^T\, V_i^{-1} \tag{17}$$

where the covariance P_i is defined as $P_i \equiv E((x_i - \hat{x}_i)(x_i - \hat{x}_i)^T)$. The difference $x_i - \hat{x}_i$ is defined as the estimation error, e_i. From the definition of P_i and the difference equation (1), (15) and (16), a difference equation for P_i is found as

$$P_i = M_i - M_i H_i^T (H_i M_i H_i^T + V_i)^{-1}\, H_i M_i \tag{18}$$

where M_i is the extrapolated covariance of the error in the estimate

$$M_i = E((x_i - \bar{x}_i)(x_i - \bar{x}_i)^T) = A_{i-1}\, P_{i-1}\, A_{i-1}^T + \Gamma_{i-1}\, W_{i-1}\, \Gamma_{i-1}^T \tag{19}$$

The error in the estimate, e_i, is independent of the measurement history Z_i and also independent of $\hat{x}_i$. The measurement residual, s_i, is a white sequence independent of the measurement history Z_{i-1}.

From the above, $p(x_{N+1}/Z_{N+1})$ is normal with conditional mean $\hat{x}_{N+1}$ and covariance P_{N+1}. Fortunately under these conditions, the expectation in (14) can be performed in closed form by completing the square (cf. {8} theorem 3, page 127) with the result

$$J_{N+1}(Z_{N+1}) = \mu\alpha_{N+1}\, \exp\,\{\tfrac{\mu}{2}\,(\sum_{i=1}^{N}\, u_i^T\, R_i u_i + \hat{x}_{N+1}^T\, \tilde{Q}_{N+1}\, \hat{x}_{N+1})\} \tag{20}$$

where

$$\tilde{Q}_{N+1} = Q_{N+1} + \mu Q_{N+1}(P_{N+1}^{-1} - \mu Q_{N+1})^{-1}\, Q_{N+1} \triangleq Q_{N+1}(I - \mu P_{N+1}\, Q_{N+1})^{-1} \tag{21}$$

The matrix identity only holds when Q_{N+1} is positive definite. Furthermore, (21) is subject to the restriction $P_{N+1}^{-1} > \mu Q_{N+1}$ and the coefficient α_{N+1} is

$$\alpha_{N+1} = (|(P_{N+1}^{-1} - \mu Q_{N+1})^{-1}|/|P_{N+1}|)^{1/2} = |I - \mu P_{N+1}\, Q_{N+1}|^{-1/2} \tag{22}$$

where $|(\cdot)|$ denotes the determinant of a matrix $(\cdot)$.

A backward step is taken to stage N by applying the recursion formula (12) to (20) as

$$J_N(Z_N) = E_{Z_N}(J_{N+1}(Z_{N+1})) \quad .$$

Since u_i is a Borel function of Z_i, it is measurable with respect to Ξ_N for $i \epsilon (1,N)$ and $J_N(Z_N)$ can be written

$$J_N(Z_N) = \mu\alpha_{N+1}\, \exp\,\{\tfrac{\mu}{2}\, \sum_{i=1}^{N}\, u_i^T\, R_i u_i\}\, E_{Z_N}\left[\exp\,\{\tfrac{\mu}{2}\, \hat{x}_{N+1}^T\, \tilde{Q}_{N+1}\, \hat{x}_{N+1}\}\right] \tag{23}$$

From (15) and (16) the expectation is taken over the measurement residual s_{N+1} which at any stage i is normal with zero mean and covariance

$$S_i = E(s_i s_i^T) = H_i M_i H_i^T + V_i \tag{24}$$

The indicated conditional expectation in (23) can be obtained in closed form by completing the square

$$J_N(Z_N) = \mu\alpha_N\, \exp\,\{\tfrac{\mu}{2}(\sum_{i=1}^{N}\, u_i^T\, R_i u_i + \bar{x}_{N+1}^T\, \underline{Q}_{N+1}\, \bar{x}_{N+1})\} \tag{25}$$

where

$$\underline{Q}_{N+1} = \tilde{Q}_{N+1} + \mu\tilde{Q}_{N+1}\, K_{N+1}(S_{N+1}^{-1} - \mu K_{N+1}^T\, \tilde{Q}_{N+1}\, K_{N+1})^{-1}\, K_{N+1}^T\, \tilde{Q}_{N+1} \tag{26}$$

subject to the restriction $S_{N+1}^{-1} > \mu K_{N+1}^T \tilde{Q}_{N+1} K_{N+1}$ and the coefficient α_N is

$$\alpha_N = \alpha_{N+1} |I - \mu K_{N+1}^T \tilde{Q}_{N+1} K_{N+1}|^{-1/2} \tag{27}$$

Finally, the extrapolated estimate is replaced by (16) in (25) giving the cost to go at stage N as

$$J_N(Z_N) = \mu\alpha_N \exp \{\frac{\mu}{2}(\sum_{i=1}^{N} u_i^T R_i u_i + (\hat{x}_N^T A_N^T + u_N^T B_N^T)\underline{Q}_{N+1}(A_N\hat{x}_N + B_N u_N))\} \tag{28}$$

Using (11) the problem of finding the minimum value of J, J*, is obtained recursively by use of the fundamental theorem in {7, page 769} which states that†

$$J^* = \underset{u_i \forall i \varepsilon [1,N]}{\text{Minimum}} \; E(J_N(Z_N)) = \underset{u_i \forall i \varepsilon [1,N-1]}{\text{Minimum}} \; E(\underset{u_N}{\text{Min}} \; J_N(Z_N)) \tag{29}$$

The minimum cost to go from stage N to N+1

$$J_N^*(Z_N) \equiv \underset{u_N}{\text{Min}} \; J_N(Z_N)$$

is obtained by introducing (28) into (29) and minimizing with respect to u_N. The resulting feedback controller is

$$u_N = -(R_N + B_N^T \underline{Q}_{N+1} B_N)^{-1} B_N^T \underline{Q}_{N+1} A_N \hat{x}_N \tag{30}$$

where u_N implies a minimum of $J_N(Z_N)$ if

$$(R_N + B_N^T \underline{Q}_{N+1} B_N) > 0$$

Substitution of (30) into (28) yields

$$J_N^*(Z_N) = \mu\alpha \exp \{\frac{\mu}{2} (\sum_{i=1}^{N-1} u_i^T R_i u_i + \hat{x}_N^T \tilde{Q}_N \hat{x}_N)\} \tag{31}$$

where

$$\tilde{Q}_N = A_N^T(\underline{Q}_{N+1} - \underline{Q}_{N+1} B_N(B_N^T \underline{Q}_{N+1} B_N + R_N)^{-1} B_N^T \underline{Q}_{N+1}) A_N \tag{32}$$

Applying this procedure to each successive backward step, the optimal control is obtained at each stage. The recursion relationships corresponding to (26) and (32) are

$$\underline{Q}_i = \tilde{Q}_i + \mu\tilde{Q}_i K_i (S_i^{-1} - \mu K_i^T \tilde{Q}_i K_i)^{-1} K_i^T \tilde{Q}_i \tag{33}$$

$$\tilde{Q}_{i-1} = A_{i-1}^T(\underline{Q}_i - \underline{Q}_i B_{i-1}(B_{i-1}^T \underline{Q}_i B_{i-1} + R_{i-1})^{-1} B_{i-1}^T \underline{Q}_i)A_{i-1} \tag{34}$$

with the terminal boundary condition given by (21) and subject to the restriction that $S_i^{-1} > \mu K_i^T \tilde{Q}_i K_i$. The optimal control is a linear function of the state estimate at each stage

$$u_i = -(R_i + B_i^T \underline{Q}_{i+1} B_i)^{-1} B_i^T \underline{Q}_{i+1} A_i \hat{x}_i \tag{35}$$

if $R_i + B_i^T \underline{Q}_{i+1} B_i > 0$. In determining the optimal control for the terminal cost problem, $\hat{x}_i$ is a sufficient statistic {10}.

The unconditional expectation of the cost is

$$J^* = \mu\alpha_1 \exp \{\frac{\mu}{2} \bar{x}_1^T \underline{Q}_1 \bar{x}_1\} \tag{36}$$

where $\bar{x}_1$ is the apriori estimate of the state and the coefficient α_1 is obtained through the recursion corresponding to (27)

$$\alpha_i = \alpha_{i+1} |I - \mu K_{i+1}^T \tilde{Q}_{i+1} K_{i+1} S_{i+1}|^{-1/2} \tag{37}$$

with terminal boundary condition given by (22).

3.2 The General Stochastic Optimal Control Problem

The purpose of this section is to generalize the solution to the terminal cost problem of the previous section to include a cost on the state at each stage. If the analysis proceeds, as in the previous section, by attempting to use the recursion formula (12), then the terminal condition on the recursion formula using the general quadratic argument (7) is

---

†Note that $J_N(Z_N)$ has been used for convenience to denote $J_N(Z_N, u_N(Z_N))$ where $u_N(Z_N)$ is now to be determined by the minimization in (29).

$$J_{N+1}(Z_{N+1}) = \mu \exp \{\frac{\mu}{2} \sum_{i=1}^{N} u_i^T R_i u_i\} E_{Z_{N+1}} (\exp \{\frac{\mu}{2} \sum_{i=1}^{N+1} x_i^T Q_i x_i\}) \tag{38}$$

where the control costs have been taken out of the expectations by the arguments leading to equation (14).

Since the expectation in (38) is conditioned on the measurement history, the smoothed estimate of the entire state history is required at each stage. If a composite vector of the state at each stage is defined as

$$X_i^T = (x_1^T , x_2^T , \dots , x_i^T) \tag{39}$$

where X_i is a $(n \cdot i)$-dimensional vector, then the general problem can be cast into the form of the terminal cost problem. Define the $((N+1) \cdot n) \times ((N+1) \cdot n)$ composite cost weighting matrix as

$$L_{N+1} = \begin{bmatrix} Q_1 & & 0 \\ & Q_2 \ddots & \\ 0 & & Q_{N+1} \end{bmatrix} \tag{40}$$

If a double sum were used to weight states at different stages in the performance index, then L_{N+1} generalizes to any non-negative definite symmetric matrix. This generalization does not affect the forthcoming analysis. The terminal cost (38) becomes

$$J_{N+1}(Z_{N+1}) = \mu \exp \{\frac{\mu}{2} \sum_{i=1}^{N} u_i^T R_i u_i\} E_{Z_{N+1}} (\exp \{\frac{\mu}{2} X_{N+1}^T L_{N+1} X_{N+1}\}) \tag{41}$$

Since the expectation in (41) is over a function of X_{N+1} given Z_{N+1}, the probability density function for X_{N+1} given Z_{N+1}, $p(X_{N+1}/Z_{N+1})$, is required. To determine $p(X_i/Z_i)$ at any stage i, the stochastic system (1) and (4) is used to develop a stochastic system in terms of X_i as

$$X_{i+1} = \tilde{A}_i X_i + \tilde{B}_i u_i + \tilde{\Gamma}_i w_i \, , \quad z = \tilde{H}_i X_i + v \tag{42}$$

where

$$\tilde{A}_i = \left[\begin{array}{c|c} \multicolumn{2}{c}{I_i} \\ \hline 0 & A_i \end{array}\right], \quad \tilde{B}_i = \begin{bmatrix} 0 \\ \hline B_i \end{bmatrix}, \quad \tilde{\Gamma}_i = \begin{bmatrix} 0 \\ \hline \Gamma_i \end{bmatrix}, \quad \tilde{H}_i^T = \begin{bmatrix} 0 \\ \hline H_i^T \end{bmatrix} \tag{43}$$

are $(i+1) \cdot n \times i \cdot n$, $(i+1) \cdot n \times p$, $(i+1) \cdot n \times m$ and $(i+1) \cdot n \times q$ matrices, respectively. Note that at each step the vector X_i increases in dimension by n. Since (42) forms a linear gaussian stochastic system, $p(X_i/Z_i)$ is normal with conditional mean $\hat{X}_i \equiv E_{Z_i}(X_i)$. If the Kalman filter gain is

$$\tilde{K}_i = \tilde{P}_i \tilde{H}_i^T V_i^{-1} \tag{44}$$

where the covariance matrix $\tilde{P}_i$ is defined as $E((X_i - \hat{X}_i)(X_i - \hat{X}_i)^T)$, then the Kalman filter equations given by (15) to (19) can be used for the smoothing equations when X_i, $\bar{X}_i$, $\hat{X}_i$, $\tilde{K}_i$, $\tilde{A}_i$, B_i, P_i, M_i replace x_i, $\bar{x}_i$, $\hat{x}_i$, K_i, A_i, B_i, P_i, M_i, respectively. At each stage P_i and M_i increase in dimension by n rows and n columns. In this way a compact scheme for performing smoothing problems in the context of the well known estimation formulation {1,2,7} is obtained. Again note the useful properties that the estimation errors $(X_i - \hat{X}_i)$ are independent of Z_i and s_i is independent of Z_{i-1}.

Since the general problem has been reformulated to parallel that of the terminal control problem, the analysis of the preceding section can be routinely applied resulting in the recursion formulas

$$\tilde{L}_{N+1} = L_{N+1}(I - \mu\tilde{P}_{N+1} L_{N+1})^{-1} \tag{45}$$

$$\underline{L}_i = \tilde{L}_i + \mu\tilde{L}_i\tilde{K}_i(S_i^{-1} - \mu\tilde{K}_i^T \tilde{L}_i\tilde{K}_i)^{-1} \tilde{K}_i^T \tilde{L}_i \tag{46}$$

$$\tilde{L}_{i-1} = \tilde{A}_{i-1}^T(\underline{L}_i - \underline{L}_i\tilde{B}_{i-1}(\tilde{B}_{i-1}^T \underline{L}_i\tilde{B}_{i-1} + R_{i-1})^{-1} \tilde{B}_{i-1}^T \underline{L}_i)\tilde{A}_{i-1} \tag{47}$$

with the restrictions that $\tilde{P}_{N+1}^{-1} - L_{N+1} > 0$ and $S_i^{-1} - \mu\tilde{K}_i^T \tilde{L}_i\tilde{K}_i > 0$. Note that the matrix identity of (21) is used in (45). The optimal control at any stage i is

$$u_i = -(R_i + \tilde{B}_i^T \underline{L}_{i+1} \tilde{B}_i)^{-1} \tilde{B}_i^T \underline{L}_{i+1} \tilde{A}_i\hat{X}_i \tag{48}$$

if $(R_i + \tilde{B}_i^T \underline{L}_{i+1} \tilde{B}_i) > 0$. Thus, in the general case, the optimal control is a linear function of the smoothed estimate of the entire state history. Note that since A_{i-1} is an $i \cdot n\text{x}(i-1) \cdot n$ matrix, the $\tilde{L}_i$ matrix decreases in size by n rows and columns at each stage along the backward recursion.

The unconditional expectation of the cost is

$$J^* = \mu\alpha_i \exp \{\frac{\mu}{2} \bar{x}_1^T \underline{L}_1 \bar{x}_1\} \tag{49}$$

where $\bar{x}_1$ is the apriori estimate of the state and the coefficient α_1 is obtained through the recursion relation

$$\alpha_{i-1} = \alpha_i |I - \mu \tilde{K}_i^T \tilde{L}_i \tilde{K}_i S_i|^{-1/2} \quad ; \quad \alpha_{N+1} = |I - \mu L_{N+1} \tilde{P}_{N+1}|^{-1/2} \tag{50}$$

3.3 No Process Noise Case

The general solution of the exponential product cost function problem yields a control which depends on the smoothed state history. If, however, there is no process noise entering the system, then the problem can be transformed to a terminal cost problem, with the addition of an extra n-dimensional state vector. The resulting control becomes a linear function of a 2n dimensional state space at each stage, rather than the entire smoothed history whose dimension increases by n at each stage.

The cost is the general form, given in (6) and (7), subject now to dynamic constraints

$$x_{i+1} = A_i x_i + B_i u_i \tag{51}$$

with only initial condition uncertainty as given in (3) and measurements as in (4). Since in (51) the state is not forced by a white noise process, the entire state history is projected to the terminal state N+1 by repeated application of (51). Assuming the A_j matrices are non-singular for all $j \varepsilon [1,N]$

$$x_i = \Phi^{-1}_{N+1,i} (x_{N+1} - \sum_{j=1}^{N} \Phi_{N+1,j+1} B_j u_j) \tag{52}$$

where the transition matrix $\Phi_{k,\ell}$ is defined as

$$\Phi_{k,\ell} = \prod_{j=\ell}^{k-1} A_j \quad ; \quad \Phi_{k,k} = I \tag{53}$$

Introducing (52) into the cost function argument ψ yields

$$\psi = x^T_{N+1} C_{N+1} x_{N+1} - 2x^T_{N+1} (\sum_{i=1}^{N} C'_i \sum_{j=1}^{N} \Phi_{N+1,j+1} B_j u_j) + \sum_{i=1}^{N} ((\sum_{j=1}^{N} \Phi_{N+1,j+1} B_j u_j)^T C'_i \sum_{k=i}^{N} \Phi_{N+1,k+1} B_k u_k))$$
$$+ \sum_{i=1}^{N} u_i^T R_i u_i \tag{54}$$

where

$$C'_i = \Phi^{-T}_{N+1,i} Q_i \Phi^{-1}_{N+1,i} \; , \; C_j = \sum_{i=1}^{j} C'_i \tag{55}$$

The multiple sums in (54) are manipulated to yield more useful expressions. Interchanging the order of summation of the double sum in the second term of (54) reduces to η_{N+1} which is the terminal value of a new n-dimensional state vector propagated as

$$\eta_{i+1} = \eta_i + C_i \Phi_{N+1,i+1} B_i u_i \quad ; \quad \eta_1 = 0 \tag{56}$$

Since the controls are Borel functions of the measurement history, η_{i+1} is a Borel function of Z_i.

For the third term in (54), let γ_{ijk} denote the argument of the triple sum. Then interchanging the order of summation of i with j

$$\sum_{i=1}^{N} \sum_{j=i}^{N} (\sum_{k=i}^{N} \gamma_{ijk}) = \sum_{j=1}^{N} \sum_{i=1}^{j} (\sum_{k=i}^{N} \gamma_{ijk}) \tag{57}$$

If the inner sum of (57) is decomposed as $\sum_{k=i}^{N} = \sum_{k=i}^{j} + \sum_{k=j+1}^{N}$, then k is interchanged with i and j in the following

$$\sum_{i=1}^{N} \sum_{j=i}^{N} \sum_{k=i}^{N} \gamma_{ijk} = \sum_{j=1}^{N} (\sum_{k=1}^{j} \sum_{i=1}^{k} \gamma_{ijk} + \sum_{k=j+1}^{N} \sum_{i=1}^{j} \gamma_{ijk}) = \sum_{j=1}^{N} \sum_{k=1}^{j} \sum_{i=1}^{k} \gamma_{ijk}$$
$$+ \sum_{j=1}^{N} (\sum_{k=j}^{N} \sum_{i=1}^{j} \gamma_{ijk} - \sum_{i=1}^{j} \gamma_{ijj}) = 2 \sum_{j=1}^{N} \sum_{k=1}^{j} \sum_{i=1}^{k} \gamma_{ijk} - \sum_{j=1}^{N} \sum_{i=1}^{j} \gamma_{ijj} \tag{58}$$

Substituting the argument of the triple sum in (54) for γ_{ijk} and defining $D_j \equiv \Phi_{N+1,j+1} B_j$, the performance index is now written in terms of a 2n state vector $(X^T_{N+1}, \eta^T_{N+1})$ as

$$J = E \left[\mu \exp \frac{\mu}{2} \left\{ (x^T_{N+1}, \eta^T_{N+1}) \begin{bmatrix} C_{N+1} & -I \\ -I & 0 \end{bmatrix} \begin{bmatrix} x_{N+1} \\ \eta_{N+1} \end{bmatrix} + \sum_{i=1}^{N} (\eta_i^T, u_i^T) \begin{bmatrix} 0 & D_i \\ D_i^T & R_i + D_i^T C_i D_i \end{bmatrix} \begin{bmatrix} \eta_i \\ u_i \end{bmatrix} \right\} \right] \tag{59}$$

Again the recursion relation (12) is used to find the optimal controller. Since η_i is a Borel function of Z_{i-1}, the cost on the control and η_i can be taken out of the expectation of the terminal cost as

$$J_{N+1}(Z_{N+1}) = \mu \exp\left\{\frac{\mu}{2}\sum_{i=1}^{N}(\eta_i^T, u_i^T)\begin{bmatrix}0 & D_i\\ D_i^T & R_i+D_i^TC_iD_i\end{bmatrix}\begin{bmatrix}\eta_i\\ u_i\end{bmatrix}\right\} \times E_{Z_{N+1}}\left[\exp\frac{\mu}{2}\left\{(x_{N+1}^T, \eta_{N+1}^T)\begin{bmatrix}C_{N+1} & -I\\ -I & 0\end{bmatrix}\begin{bmatrix}x_{N+1}\\ \eta_{N+1}\end{bmatrix}\right\}\right] \tag{60}$$

As in Section 3.1 the expectation over x_{N+1} given Z_{N+1} can be obtained in closed form as

$$J_{N+1}(Z_{N+1}) = \mu\alpha_{N+1} \exp\left\{\frac{\mu}{2}\left(\sum_{i=1}^{N}(\hat{y}_i^T, u_i^T)\begin{bmatrix}0 & \tilde{\tilde{D}}_i\\ \tilde{\tilde{D}}_i^T & \underline{R}_i\end{bmatrix}\begin{bmatrix}\hat{y}_i\\ u_i\end{bmatrix} + \hat{y}_{N+1}^T\,\tilde{\tilde{C}}_{N+1}\,\hat{y}_{N+1}\right)\right\} \tag{61}$$

where for convenience

$$\hat{y}_i^T = (\hat{x}_i^T, \eta_i^T)\ ,\quad \tilde{\tilde{D}}_i^T = (0, D_i^T)\ ,\quad \underline{R}_i = R_i + D_i^T\,C_iD_i \tag{62}$$

The estimation equations are now

$$\hat{y}_{i+1} = \bar{y}_{i+1} + \tilde{\tilde{K}}_{i+1}\,s_{i+1},\quad \bar{y}_{i+1} = (\bar{x}_{i+1}^T, \eta_{i+1}^T) = \tilde{\tilde{A}}_i\hat{y}_i + \tilde{\tilde{B}}_iu_i \tag{63}$$

where

$$\tilde{\tilde{K}}_{i+1} = \begin{bmatrix}K_{i+1}\\ 0\end{bmatrix},\quad \tilde{\tilde{A}}_i = \begin{bmatrix}A_i & 0\\ 0 & I\end{bmatrix},\quad \tilde{\tilde{B}}_i = \begin{bmatrix}B_i\\ C_iD_i\end{bmatrix} \tag{64}$$

and where s_{i+1} is the measurement residual defined in (16) having zero mean and variance S_{i+1} (24) and K_{i+1} is the Kalman gain defined by (17) to (19).

In this form (60) the recursion relation (12) can be used just as in Section 3.1 to obtain recursion formulas and an optimal control law as

$$\tilde{\tilde{C}}_{n+1} = \begin{bmatrix}C_{N+1}(I - \mu P_{N+1}\,C_{N+1})^{-1} & -(I - \mu C_{N+1}\,P_{N+1})^{-1}\\ -(I - \mu P_{N+1}\,C_{N+1})^{-1} & \mu(P_{N+1}^{-1} - \mu C_{N+1})^{-1}\end{bmatrix} \tag{65}$$

$$\underline{C}_i = \tilde{\tilde{C}}_i + \mu\tilde{\tilde{C}}_i\tilde{\tilde{K}}_i(S_i^{-1} - \mu\tilde{\tilde{K}}_i^T\,\tilde{\tilde{C}}_i\tilde{\tilde{K}}_i)^{-1}\,\tilde{\tilde{K}}_i^T\,\tilde{\tilde{C}}_i \tag{66}$$

$$\tilde{\tilde{C}}_{i-1} = \tilde{\tilde{A}}_{i-1}^T\,\underline{C}_i\tilde{\tilde{A}}_{i-1} - (\tilde{\tilde{A}}_{i-1}^T\,\underline{C}_i\tilde{\tilde{B}}_{i-1} + \tilde{\tilde{D}}_{i-1})(\underline{R}_{i-1} + \tilde{\tilde{B}}_{i-1}^T\,\underline{C}_i\tilde{\tilde{B}}_{i-1})^{-1}(\tilde{\tilde{D}}_{i-1}^T + \tilde{\tilde{B}}_{i-1}^T\,\underline{C}_i\tilde{\tilde{A}}_{i-1}) \tag{67}$$

with the restriction that $P_{N+1}^{-1} - \mu C_{N+1} > 0$ and $S_i^{-1} - \mu\tilde{\tilde{K}}_i^T\,\tilde{\tilde{C}}_i\tilde{\tilde{K}}_i > 0$. Note that in (65) the matrix identity of (21) is used. The optimal control is a linear function of the new 2n dimensional state $\hat{y}_i^T = (\hat{x}_i^T, \eta_i^T)$

$$u_i = -(\underline{R}_i + \tilde{\tilde{B}}_i^T\,\underline{C}_{i+1}\,\tilde{\tilde{B}}_i)^{-1}(\tilde{\tilde{D}}_i^T + \tilde{\tilde{B}}_i^T\,\underline{C}_{i+1}\,\tilde{\tilde{A}}_i)\hat{y}_i \tag{68}$$

if $R_i + \tilde{\tilde{B}}_i^T\,\underline{C}_{i+1}\,\tilde{\tilde{B}}_i > 0$ where the additional state is the weighted sum of past values of control. Unlike the general case the control gains are independent of the number of stages. Furthermore, the unconditional expectation of the cost is

$$J^* = \mu\alpha_1 \exp\frac{\mu}{2}(\bar{x}_1^T, 0)\underline{C}_i\begin{bmatrix}\bar{x}_1\\ 0\end{bmatrix} \tag{69}$$

where α_1 is obtained by the recursion relation

$$\alpha_{i-1} = \alpha_i|I - \mu\tilde{\tilde{K}}_i^T\,\tilde{\tilde{C}}_i\tilde{\tilde{K}}_iS_i|^{-1/2}\ ,\quad \alpha_{N+1} = |I - \mu C_{N+1}\,P_{N+1}|^{-1/2} \tag{70}$$

4. ONE STAGE SCALAR EXAMPLE

The following one stage scalar example illustrates the increased flexibility of controller design using the exponential performance index. The problem is to minimize

$$J = \mu E(\exp \mu\psi/2)\ ,\quad \psi = u_1^2\,R_1 + x_2^2\,Q_2 \tag{71}$$

subject to the stochastic scalar equation

$$x_2 = x_1 + u_1 + w_1 \tag{72}$$

with measurement sequence

$$z_1 = x_1 + v_1 \; , \quad z_2 = x_2 + v_2 \tag{73}$$

where x_1, w_1, v_1, and v_2 are uncorrelated with zero mean and variances M_1, W_1, V_1, and V_2, respectively having a value of one. Also the values of R_1 and Q_2 are taken as one. First, the estimation error variances are evaluated. From (18) using the a priori error variance M_1, the posteriori error variance at stage one is $P_1 = 1/2$. From (23) the a priori error variance at stage two is $M_2 = 3/2$. Again using (18) and M_2 the posteriori error variance is $P_2 = 3/5$. The Kalman gains at stage one and two are $K_1 = 1/2$ and $K_2 = 3/5$, respectively. After the variance is evaluated in the forward direction the feedback gain sequence as a function of μ is evaluated backwards starting at (21) where $\tilde{Q}_2 = 5/(5 - 3\mu)$. Using Q_2, $K_2 = 3/5$, and $S_2 = 5/2$ evaluated from (24), $\underline{Q}_2 = 2/(2 - 3\mu)$ is evaluated from (26) where μ is restricted to be less than 2/3. Finally the control gain given by (30) is $\Lambda_1 = -2/(4 - 3\mu)$.

The performance of the controller for various values of μ is measured by evaluating the first and second moments of $\psi/2$ given as

$$E(\psi)/2 = \bar{\psi}/2 = 1 + (\Lambda_1 + \Lambda_1^2)/2 \; , \quad E((\psi - \bar{\psi})^2)/4 = 0.5\Lambda_1^4 + \Lambda_1^3 + 1.75\Lambda_1^2 + 2\Lambda_1 + 2 \tag{74}$$

It is easily seen that the minimum value of $\bar{\psi}/2$ occurs when $\mu = 0$. For $\mu = 0, 1/3, -1/3$ the values of $\bar{\psi}/2$ are 0.875, 0.89, and 0.88, respectively, and the values of the variance, $E((\psi-\psi)^2)/4$ are 1.35, 1.25, and 1.43 respectively. For the positive exponential case, a 1.7% increase in the mean is traded for a 7.4% decrease in the variance over that of the LQG solution. In the negative exponential case both the mean and variance increase over that of the LQG problem. However, if the control effort as given by $E(u_1^2)/2$ is calculated for $\mu = 0, 1/3, -1/3$ as 0.0625, 0.11, 0.04, respectively, then the savings in control effort using the negative exponential is 0.0225 obtained at the expense of an increase in mean of only 0.005 and 6.3% in variance. This savings in control effort occurs because the control gains decrease from that of the LQG problem in proportion to the uncertainty in the estimation of the state. Thereby, the controller hedges against large control effort when the information is poor. On the other hand, the positive exponential solution in order to control the variance uses considerable control effort; 0.11 in comparison to 0.0625 for the LQG problem. Furthermore, the smallest variance is achieved for $\mu = 0.65$ for which a 10.4% decrease in variance is obtained by an increase in the mean to 0.93 (a 6.3% increase) and an increase in control effort to 0.174.

5. DISCRETE TO CONTINUOUS LIMIT

In continuous time, the optimization problem takes the form of

$$\underset{u(\cdot,\ \cdot)}{\text{minimum}}\ E(\mu \exp \frac{\mu}{2} (\int_{t_o}^{t_f} (x^T Qx + u^T Ru)dt + x^T(t_f)Q_f x(t_f))) \tag{75}$$

subject to the Ito stochastic system {7}

$$dx = (Ax + Bu)dt + \Gamma dw \; ; \quad dz = Hxdt + dv \tag{76}$$

where A is a n × n matrix, B is a n × p matrix, Γ is a n × m matrix and H is a q × n matrix, where, for notational simplicity, the time dependence of the variables has been suppressed, and where w and v are gaussian brownian motion processes with zero mean and satisfying

$$E(w(t)w^T(\tau)) = \int_{t_o}^{\min(t,\tau)} W(s)ds \; , \quad E(v(t)v^T(\tau)) = \int_{t_o}^{\min(t,\tau)} V(s)ds \tag{77}$$

The initial state x_o is a normally distributed vector independent of w and v.

The solution to this continuous problem is obtained by formally taking the limit of the discrete results of Section 3.1 and 3.3 by letting the number of stages increase while the time between stages, $t_{i+1} - t_i = \Delta$, tend to zero. The parameters of the continuous stochastic system are first related to those of the discrete system (1) and (4). The associated Ito integral form of (76) may be written as*

$$x(t_i + \Delta) = x(t_i) + (A(t_i)x(t_i) + B(t_i)u(t_i))\Delta + \Gamma(t_i) \int_{t_i}^{t_i+\Delta} dw + C_1(\Delta) \tag{78}$$

$$z(t_i + \Delta) - z(t_i) = H(t_i)x(t_i)\Delta + \int_{t_i}^{t_i+\Delta} dv + C_2(\Delta) \tag{79}$$

where $C_1(\ )$ and $C_2(\ )$ represent terms of higher order than Δ (i.e., $\lim C_i(\Delta)/\Delta \to 0$). For sufficiently small Δ and where the interval of integration in (77) is Δ, the following identifications are made:

---

*It can be shown rigorously that the limit of the difference equations approaches the Ito form (76) in finite distribution using a result in {11, page 155}.

$$x(t_i) \triangleq x_i \ , \quad \int_{t_i}^{t_i+\Delta} dw \triangleq w_i \ , \quad \int_{t_i}^{t_i+\Delta} dv \triangleq v_i \ , \quad H(t_i) \triangleq H_i \ ,$$

$$(I + A\Delta) \triangleq A_i \ , \quad B\Delta \triangleq B_i \ , \quad \Gamma(t_i) \triangleq \Gamma_i \ , \quad K\Delta \triangleq K_i \ , \tag{80}$$

$$W\Delta \triangleq W_i \ , \quad V/\Delta \triangleq V_i \ , \quad R\Delta \triangleq R_i \ , \quad Q\Delta \triangleq Q_i$$

Since the recursion relations derived in Section 3.1 and 3.3 are deterministic difference equations, the formal limit can be rigorously justified.

5.1 Continuous Time Terminal Cost Problem

The results of the terminal cost problem discussed in Section 3.1 can be written in continuous time by first substituting (80) into (33) and (34). If $\tilde{Q}_i$ and $\tilde{Q}_{i-1}$ are associated with $\tilde{Q}(t_i)$ and $\tilde{Q}(t_i - \Delta)$, if $\underline{Q}_i$ is eliminated, and if Δ goes to zero, then in the limit the Riccati equation results

$$\lim_{\Delta\to 0} (\tilde{Q}(t_i - \Delta) - \tilde{Q}(t_i))/\Delta \to -\dot{\tilde{Q}} = A^T \tilde{Q} + \tilde{Q}A - \tilde{Q}(BR^{-1} B^T - \mu KVK^T)\tilde{Q} \tag{81}$$

where K is the continuous time Kalman filter, $K = PH^TV^{-1}$ and where P, the continuous time covariance of the error in the estimate, is propagated by the Riccati equation {cf. 2, 12}. The terminal boundary condition for (81) $\tilde{Q}_f = \tilde{Q}_{N+1}$ is found by replacing Q_{N+1} by Q_f in (21). The Riccati equation (81) and its boundary condition are subject to the conditions $(BR^{-1}B^T-\gamma KVK^T)>0$, $P^{-1}(t_f)-\gamma Q_f>0$; $t\varepsilon(t_o, t_f)$.

Using (80) the recursion relation (37) for $\alpha_i = \alpha(t_i)$ with terminal boundary condition from (22) in the limit is the linear scalar equation

$$-\dot{\alpha} = \alpha\mu \operatorname{tr}(VK^T \tilde{Q}K)/2 \ , \quad \alpha(t_f) = |(I - \mu P(t_f)Q_f)^{-1}|^{1/2} \tag{82}$$

where tr denotes the trace operation. Finally, the continuous time control vector obtained from (35) is

$$u = -R^{-1} B\tilde{Q}\hat{x} \tag{83}$$

where $\hat{x}$ is propagated by the continuous form of the Kalman estimator {12}. Note that the convexity condition for the continuous time problem is $R > 0$.

The effect of the quality of the information as represented by P is seen to strongly influence the feedback gains (83) through the calculation of Q. Note that if μ becomes arbitrarily small the feedback gain of (83) and Riccati equation (81) and its boundary condition reduce to that of the standard LQG problem {2}.

5.2 Continuous Form of the 2n-Dimensional Controller

Although in the general case the controller depends in the limit on an infinite dimensional vector, Section 3.3 showed that for the no process noise case a 2n-dimensional controller results. If (80) is used in the coefficients defined in (62) and (64), then

$$\tilde{\tilde{A}}_i \sim \begin{bmatrix} I+A\Delta & 0 \\ 0 & I \end{bmatrix} = I + \tilde{\tilde{A}}\Delta \ , \quad \tilde{\tilde{B}}_i \sim \begin{bmatrix} B \\ CB \end{bmatrix} \Delta = \tilde{\tilde{B}}\Delta$$

$$\tilde{\tilde{K}}_i = \begin{bmatrix} K \\ 0 \end{bmatrix} \Delta = \tilde{\tilde{K}}\Delta \ , \quad \tilde{\tilde{D}}_i \sim \begin{bmatrix} 0 \\ B \end{bmatrix} \Delta = \tilde{\tilde{D}}\Delta \tag{84}$$

where C_i defined in (55) in the limit becomes

$$C(t) = \lim_{\substack{\Delta\to 0 \\ i\to\infty \\ i\Delta\to t}} \sum_{i=1}^{i} \Phi^{-T}_{N+1,i} \ Q_i \Phi^{-1}_{N+1,i} = \int_{t_o}^{t} \Phi^{-T}(t_f, \tau)Q(\tau)\Phi^{-1}(t_f, \tau)d\tau \tag{85}$$

and is obtained by a quadrature integration where, using the identity

$$\Phi^{-1}(t_f, t) \equiv \Phi(t, t_f) \ , \quad d\Phi(t, t_f)/dt = -\Phi(t, t_f)A \ ; \quad \Phi(t_f, t_f) = I \tag{86}$$

Note that at the final stage the terminal weighting adds to (85) as

$$C(t_f) = Q_f + C(t_{\bar{f}}) \tag{87}$$

where $C(t_{\bar{f}})$ is (85) evaluated at $t = t_f$.

Inserting (84) and (85) into the difference equation (66) and (67) the continuous time 2n×2n matrix Riccati equation results

$$-\dot{\tilde{\tilde{C}}} = \tilde{\tilde{A}}^T \tilde{\tilde{C}} + \tilde{\tilde{C}}\tilde{\tilde{A}} - (\tilde{\tilde{C}}\tilde{\tilde{B}} + \tilde{\tilde{D}})R^{-1} (\tilde{\tilde{D}}^T + \tilde{\tilde{B}}^T \tilde{\tilde{C}}) + \mu\tilde{\tilde{C}}\tilde{\tilde{K}}^T V\tilde{\tilde{K}}\tilde{\tilde{C}} \tag{88}$$

where the terminal boundary condition using (87) to replace C_{N+1} by $C(t_f)$ is given by (65). Furthermore, the recursion relation (70) for the coefficient $\alpha_i = \alpha(t_i)$ in continuous time is

$$-\dot{\alpha} = \alpha\mu\,\mathrm{tr}(V\tilde{\tilde{K}}^T\,\tilde{\tilde{C}}\tilde{\tilde{K}})/2\ , \quad \alpha(t_f) = (|(1 - \mu P(t_f)C(t_f))^{-1}|)^{1/2} \tag{89}$$

The integrations to be made are sequenced as; integrate $\Phi(t, t_f)$ backwards, P and C forward, and $\tilde{\tilde{C}}$ and α backwards. Once $\tilde{\tilde{C}}$ is calculated the control law (68) for continuous time is

$$u = -R^{-1}(\tilde{\tilde{D}}^T + \tilde{\tilde{B}}^T\,\tilde{\tilde{C}})\hat{y} \tag{90}$$

where the 2n vector $\hat{y} = (\hat{x}^T, \eta^T)$ is propagated in part by the continuous form of the estimator equation {2,12} and in part by

$$\dot{\eta} = C(t)\Phi(t_f, t)B(t)u(t) \tag{91}$$

In this new controller the weighted sum of past values of the control are to be remembered. This vector is similar to integral compensation in classical control.

6. CONCLUSIONS

The solution to the general formulation of the stochastic control problem using an exponential cost function results in a linear controller whose gain operates upon the entire smoothed state history. Although implementation of this controller becomes increasingly more awkward with the number of stages, the structure of the controller is quite simple lending itself possibly to less awkward approximate controllers. Even so, the certainty equivalence principle and separation theorem do not hold as in the very special LQG problem. In all the controllers developed here, the feedback gains are explicitly dependent upon the quality of the information. That is, the control gains in the general problem depend upon the covariance of the error vector of the entire state history. However, in the two special cases ($Q_i = 0$ $i\varepsilon(1,N)$ and $W_i = 0$ $i\varepsilon(1,N)$) only the covariance of the error in the current dynamical state is used in the gain calculations. This simplification and the dependence of the controller on a fix dimensional state vector independent of the number of stages makes implementation as easy as in the LQG case. Note that the LQG problem is a special case of this more general stochastic controller.

A simple one stage scalar example illustrates the increase flexibility in controller design using the exponential performance index which allows a degree of shaping of the density function for the quadratic form. Using a positive exponential function a compromise is obtained between increasing the mean and decreasing the variance of the quadratic form away from that of the LQG solution. For the negative exponential both the mean and the variance increase. However, the control effort was reduced for only a slight increase in the mean. One important aspect of this control law is that the control gains decrease proportionately to the quality of the information as measured by the estimation error variance. Therefore the gains "hedge" against poor information. This intuitive notion is lacking in the LQG problem.

The definition of separation given by Wohman {3} is too narrow since the controller only depends upon the current state estimate. In the no process noise case of Section 3.3 the controller depends upon the state estimate generated by a Kalman estimator and an additional n-dimensional vector. Although this does not fit Wohman's definition, this controller does fit into the general class described by Striebel {10}. Striebel defines an "information" statistic which for a linear system with no process noise and exponential cost function is equivalent to $\hat{y}$ (cf. {10}, theorem 3). Furthermore, the addition n-vector in $\hat{y}$ is reminiscent of integral compensation in classical control theory.

REFERENCES

1. Special Issue on the Linear-Quadratic-Gaussian Estimation and Control Problem, IEEE Trans. on Auto. Cont., Athens, M., Ed., Dec. 1971, Vol. AC-16, No. 6.
2. Bryson, A.E., Jr., and Ho, Y.C., Applied Optimal Control, Waltham, Mass., Blaisdell, 1969.
3. Wonham, W.M., "On the Separation Theorem of Stochastic Control", SIAM J. Control, Vol.6, No. 2, 1968.
4. Jacobson, D.H., "Optimal Stochastic Linear Systems with Exponential Performance Criteria and their Relation to Deterministic Differential Games", IEEE Trans. on Auto. Cont., Vol. AC-18, No. 2, April 1973.
5. Deyst, J.J., "Optimal Control in the Presence of Measurement Uncertainties", Sc.D. Thesis, MIT, 1969.
6. Wong, E., Stochastic Processes Information and Dynamical Systems, McGraw-Hill, New York,1971.
7. Kalman, R.E., "A New Approach to Linear Filtering and Prediction Problems", Trans. ASME, J. Basic Eng., Vol. 82, pp. 34-45, March, 1961.
8. Brockett, R., Finite Dimensional Linear Systems, New York: Wiley, 1970.
9. Meier, L.III, Lawson, R.E., and Tether, A.J., "Dynamic Programming for Stochastic Control of Discrete Systems", IEEE Trans. on Auto. Cont., Vol. AC-16, No. 6, Dec. 1971.
10. Striebel, C., "Sufficient Statistics in the Optimum Control of Stochastic Systems", J. Math. Anal. Appl., Vol. 12, 1965.
11. Skorokhod, A.V., Studies in the Theory of Random Processes, Addison-Wesley, 1965.
12. Kalman, R.E. and Bucy, R.S., "New Results in Linear Filtering and Prediction Theory", Trans. ASME J. Basic Eng., Vol. 83, pp. 95-107, Dec. 1961.

ON MINIMAX PARAMETER ESTIMATION UNDER EXPONENTIATED GENERALIZED QUADRATIC LOSS

DAVID STELTE and MAX MINTZ
Coordinated Science Laboratory and
Department of Electrical Engineering
University of Illinois
Urbana, Illinois 61801

1. INTRODUCTION

Two minimax parameter estimation problems are posed as follows:

Let z be a single observation of a scalar random variable where

$$z = \theta + v. \tag{1}$$

Let v be a scalar random variable, where $v \in N[0,\sigma^2]$, σ^2 is a known constant, and θ is an unknown element of R.

We wish to determine a minimax estimate, $\delta^*(z)$, for θ with respect to two exponentiated generalized quadratic loss functions,

$$L^+(\delta,\theta) = e^{c(\delta(z)-\theta)^2 - k\theta^2}, \tag{2}$$

and

$$L^-(\delta,\theta) = -e^{-c(\delta(z)-\theta)^2 + k\theta^2}, \tag{3}$$

where c > 0, and k > 0 are two weighting constants.

The Exponentiated Generalized Quadratic Loss Functions $L^+(\delta,\theta)$, and $L^-(\delta,\theta)$, and Their Relation to the Generalized Quadratic Loss Function:

By a generalized quadratic loss function, we mean a loss function of the form

$$L_o(\delta,\theta) = c(\delta-\theta)^2 - k\theta^2, \tag{4}$$

where c,k > 0.

The second term in the generalized quadratic loss function places a quadratic cost on Nature's choice of θ. This loss function, as compared with the standard quadratic loss function, forces nature to compromise between a gain due to estimation error $[c(\delta-\theta)^2]$ and a loss which is equal to $[-k\theta^2]$.

The relationship between $L^+(\delta,\theta)$ and $L_o(\delta,\theta)$, and the relationship between $L^-(\delta,\theta)$ and $L_o(\delta,\theta)$ are:

$$L^+(\delta,\theta) = e^{L_o(\delta,\theta)}$$

and

$$L^-(\delta,\theta) = -e^{-L_o(\delta,\theta)}.$$

The solution to the related minimax estimation problem with a generalized quadratic loss function has been considered previously by

Basar and Mintz [1]. In this case, the minimax decision rule has been shown to be a linear rule, which is Bayes with respect to a certain normal prior distribution for θ.

The pair of minimax parameter estimation problems which we consider here is related, both by structure and solution, to a pair of Bayesian stochastic control problems which were considered recently by Jacobson in [2].

Loss Function Motivation - Some Salient Characteristics of $L^+(\delta,\theta)$ and $L^-(\delta,\theta)$:

The formulation of the preceding minimax estimation problems is motivated by the desire to investigate the canonical questions of existence, structure, and behavior of minimax decision rules for loss functions which force Nature to compromise between a gain due to estimation error and a loss due to the choice of θ. In particular, we wish to consider loss functions which exhibit the following structural behavior:

$$L(\delta,\theta) = \alpha f(\delta-\theta)g(\theta),$$

where $f:R \to R$, is symmetric about zero, and is monotone nondecreasing for positive arguments; $g:R \to R$, is symmetric about zero, and is monotone non-increasing for positive arguments; and $\alpha = \pm 1$. The choice of $\alpha = \pm 1$ is made to insure that $\alpha f(\delta-\theta)g(\theta)$ is monotone nondecreasing in $|\delta-\theta|$ for fixed θ, and monotone non-increasing in $|\theta|$ for fixed $(\delta-\theta)$. A loss function with these characteristics forces Nature to compromise between a gain due to estimation error [the factor $f(\cdot)$] and a loss due to the choice of θ [the factor $g(\cdot)$]. The loss functions $L^+(\delta,\theta)$ and $L^-(\delta,\theta)$ have these desired properties.

In comparing $L^+(\delta,\theta)$ to $L_o(\delta,\theta)$ we see that $L^+(\delta,\theta)$ accentuates the Statistician's loss [$L_o(\delta,\theta) > 0$] while discounting Nature's loss [$L_o(\delta,\theta) < 0$]. In contrast, $L^-(\delta,\theta)$ discounts the Statistician's loss [$L_o(\delta,\theta) > 0$] while accentuating Nature's loss [$L_o(\delta,\theta) < 0$].

2. MAIN RESULTS

The conditional risk of δ given θ will be denoted by $r[\delta,\theta]$ and is identical to $E[L(\delta,\theta)|\theta]$.

Definition: A decision rule $\delta^*(z)$ is said to be minimax if for all rules $\pi(z)$,

$$\sup_{\theta\in R} r[\delta^*,\theta] \leq \sup_{\theta\in R} r[\pi,\theta].$$

Definition: A probability distribution $\lambda^*(\theta)$ is said to be a least favorable prior distribution if for all probability measures $\mu(\theta)$,

$$\inf_{\delta}\int r[\delta,\theta]d\lambda^*(\theta) \geq \inf_{\delta}\int r[\delta,\theta]d\mu(\theta).$$

Minimax estimates $\delta^*(z)$ for θ and least favorable prior probability distributions $\lambda^*(\theta)$ for θ with respect to loss functions (2) and (3) are given by Theorems 1 and 2.

Theorem 1: Let $L^{+}(\delta,\theta)$ be defined by (2).

i) a) An admissible minimax estimate $\delta^{*}_{+}(z)$ for θ is given by

$$\delta^{*}_{+}(z) = 0 \quad ; \; k \geq c,$$

$$\delta^{*}_{+}(z) = \frac{c - \sqrt{kc(1 + 2k\sigma^{2} - 2c\sigma^{2})}}{c(2k\sigma^{2} + 1)} z; \; k < c < k + \frac{1}{2\sigma^{2}} .$$

b) An admissible minimax estimate $\delta^{*}_{+}(z)$ for θ when $c = k + \frac{1}{2\sigma^{2}}$ is

$$\delta^{*}_{+}(z) = \frac{1}{1 + 2k\sigma^{2}} z .$$

c) If $c > k + \frac{1}{2\sigma^{2}}$, any decision rule $\delta(z)$ has unbounded conditional risk and is therefore minimax.

ii) A least favorable prior distribution $\lambda^{*}_{+}(\theta)$ for θ is

$$\theta = 0 \text{ WP1} \quad ; \quad k \geq c,$$

$$\theta \in N[0,r^{2}] \quad ; \quad k < c < k + \frac{1}{2\sigma^{2}} ,$$

where

$$r^{2} = \frac{a\sigma^{2}}{1 - a - 2\sigma^{2}ka} ,$$

and a is the z coefficient of the linear decision rule in i),

$$a = \frac{c - \sqrt{kc(1 + 2k\sigma^{2} - 2c\sigma^{2})}}{2kc\sigma^{2} + c} .$$

Theorem 2: Let $L^{-}(\delta,\theta)$ be defined by (3).

i) An admissible minimax estimate $\delta^{*}_{-}(z)$ for θ is given by

$$\delta^{*}_{-}(z) = 0 \quad ; \quad k \geq c,$$

$$\delta^{*}_{-}(z) = \frac{-c + \sqrt{kc(2c\sigma^{2} - 2k\sigma^{2} + 1)}}{c(2k\sigma^{2} - 1)} z; \; k < c, \; 2k\sigma^{2} \neq 1$$

$$\delta^{*}_{-}(z) = \frac{c-k}{2c} z; \quad k < c, \; 2k\sigma^{2} = 1.$$

ii) A least favorable prior probability distribution $\lambda^{*}_{-}(\theta)$ is

$$\theta = 0 \text{ WP1} \quad ; \quad k \geq c,$$

$$\theta \in N[0,r^{2}] \quad ; \quad k < c,$$

where

$$r^2 = \frac{a\sigma^2}{1 - a + 2\sigma^2 ka},$$

and a is the z coefficient of the linear decision rule in i),

$$a = \frac{-c + \sqrt{kc(2c\sigma^2 - 2k\sigma^2 + 1)}}{c(2k\sigma^2 - 1)} ; \quad 2k\sigma^2 \neq 1,$$

$$a = \frac{c-k}{2c} ; \quad 2k\sigma^2 = 1.$$

Most of Theorem 1 is proven in the Appendix. Complete proofs of Theorems 1 and 2 can be found in [6].

3. FURTHER DISCUSSION

In this paper we have considered the existence and structure of the solutions of two minimax parameter estimation problems which are characterized by the exponentiated generalized quadratic loss functions $L^+(\delta,\theta)$ and $L^-(\delta,\theta)$. In both cases, minimax decision rules and (where applicable) least favorable prior distributions for the unknown parameters have been determined. In the following table (Table I) we compare the solutions of the two minimax parameter estimation problems considered in this paper with the analogous results of [1] for the generalized quadratic loss function.

An analysis of Table I shows the following:

(i) When $c \leq k$, then $a = 0$, and $r^2 = 0$ in all cases.

(ii) When $c > k$, the coefficient a_o is independent of σ^2, whereas a_+ and a_- depend on σ^2.

(iii) When $c > k + 1/2\sigma^2$, the minimax risk associated with $L^+(\delta,\theta)$ is $+\infty$, whereas the minimax risk associated with $L^-(\delta,\theta)$ and $L_o(\delta,\theta)$ is finite for all admissible values of c, k, and σ^2.

The pair of minimax parameter estimation problems which we have considered in this paper is related, both by structure and solution, to a pair of Bayesian stochastic control problems with perfect state observations which was considered recently by Jacobson in [2]. The problem statement of [2] includes a linear plant driven by white Gaussian noise and a pair of exponentiated quadratic cost functions. If we were to recast our minimax estimation problems by (i) setting k equal to zero, (ii) letting $\theta \in N[0,r^2]$ where r^2 is given, and (iii) seeking the related Bayes decision rules for the loss functions $L^+(\delta,\theta)$ and $L^-(\delta,\theta)$, we would have a pair of Bayesian optimization problems of the variety considered in [2].

It is shown in [2] that the optimal stochastic control laws are linear control laws with gains which depend on the covariance of the plant noise. This result is in contrast to the standard LQG control problem, where the optimal (Bayes) control law is a linear control law which is functionally independent of the plant noise covariance. It is further shown in [2] that under certain circumstances there will be no control law with a finite cost for the control problem with the exponentiated quadratic

Table I

c, k, σ^2	$\delta_o^*(z) = a_o z$ $\lambda_o^*(\theta) = N[0, r_o^2]$ (see [1].)	$\delta_-^*(z) = a_- z$ $\lambda_-^*(\theta) = N[0, r_-^2]$	$\delta_+^*(z) = a_+ z$ $\lambda_+^*(\theta) = N[0, r_+^2]$
Case I. $c \leq k$	$a_o = 0$ $r_o^2 = 0$	$a_- = 0$ $r_-^2 = 0$	$a_+ = 0$ $r_+^2 = 0$
Case IIa. $k < c < k + 1/2\sigma^2$, $2k\sigma^2 \neq 1$	$a_o = 1 - (k/c)^{1/2}$ $r_o^2 = [(c/k)^{1/2} - 1]\sigma^2$	$a_- = \frac{-c + [kc(2c\sigma^2 - 2k\sigma^2 + 1)]^{1/2}}{c(2k\sigma^2 - 1)}$ $r_-^2 = \frac{(a_-)\sigma^2}{1 - (a_-) + 2\sigma^2 k(a_-)}$	$a_+ = \frac{c - [kc(2k\sigma^2 - 2c\sigma^2 + 1)]^{1/2}}{c(2k\sigma^2 + 1)}$ $r_+^2 = \frac{(a_+)\sigma^2}{1 - (a_+) - 2\sigma^2 k(a_+)}$
Case IIb. $k < c < k + 1/2\sigma^2$, $2k\sigma^2 = 1$	$\left.\begin{array}{l} a_o \\ r_o^2 \end{array}\right\}$ Same as case IIa.	$a_- = \frac{c-k}{2c}$ $r_-^2 = \frac{(a_-)\sigma^2}{1 - (a_-) + 2\sigma^2 k(a_-)} = (a_-)\sigma^2$	$\left.\begin{array}{l} a_+ \\ r_+^2 \end{array}\right\}$ Same as case IIa.
Case III. $c = k + 1/2\sigma^2$	$\left.\begin{array}{l} a_o \\ r_o^2 \end{array}\right\}$ Same as case IIa.	$\left.\begin{array}{l} a_- \\ r_-^2 \end{array}\right\}$ See case IIa if $2k\sigma^2 \neq 1$, otherwise see case IIb.	$a_+ = 1/(1 + 2k\sigma^2)$ $r_+^2 = +\infty$ (λ_+^* is not defined)
Case IV. $c > k + 1/2\sigma^2$	$\left.\begin{array}{l} a_o \\ r_o^2 \end{array}\right\}$ Same as case IIa.	$\left.\begin{array}{l} a_- \\ r_-^2 \end{array}\right\}$ See case IIa if $2k\sigma^2 \neq 1$, otherwise see case IIb.	All decision rules have unbounded conditional risk functions.

cost function with the positive (or nonnegative) quadratic forms. This result is related to our present problem, where if $c > k + 1/2\sigma^2$, then all decision rules under $L^+(\delta,\theta)$ have unbounded conditional risk.

It should be noted that the results in [2] are Bayes solutions to problems with complete statistical descriptions, whereas the minimax estimation problems considered in this paper are statistical decision problems with unknown parameters which have no prior statistics. We note however that we have employed Bayesian techniques to determine the indicated minimax decision rules. Specifically, we have used the method of least favorable prior distributions. The details of this approach are contained in the following Appendix.

APPENDIX

We begin by stating several lemmata which will be useful in proving Theorem 1.

Lemma 1 (Wald): Let λ denote a prior probability measure for θ. Further, let $\delta_\lambda(z)$ denote a Bayes estimate for θ with respect to λ, and an arbitrary risk function. Assume that $\max_{\theta \in R} r[\delta_\lambda,\theta]$ exists, and denote its value by $\overline{r}_\lambda$. Let

$$\tau_\lambda = \{\theta : r[\delta_\lambda,\theta] = \overline{r}_\lambda\}.$$

If $\lambda[\tau_\lambda] = 1$, then the Bayes estimate $\delta_\lambda(z)$ is a minimax estimate for θ, and λ is a least favorable prior probability measure for θ.

A proof of Lemma 1 can be found in Lehmann [3].

Lemma 2 (Sherman): Let $p(\theta|z)$ be a conditional (posterior) probability density for θ, symmetric about its conditional mean $\mu = E[\theta|z]$. Let $L(\delta,\theta) = L(e)$, where $e = \delta(z) - \theta$, and assume $L(e)$ is a strictly convex function in e, symmetric about $e = 0$. Then the Bayes rule $\delta_\lambda(z)$ is unique up to equivalence, and identical to the conditional mean μ.

A proof of Lemma 2 can be found in Viterbi [4].

Lemma 3 (Wald): Let $\lambda_1,\lambda_2,\ldots,$ be a sequence of distributions for θ. Let δ_{λ_n} be a Bayes rule with respect to λ_n, and let ρ_n be the corresponding Bayes risk. If $\lim_{n\to\infty} \rho_n = r_c$ and δ is any estimate having constant conditional risk r_c, then δ is minimax.

A proof of Lemma 3 can be found in Lehmann [3], p. 4-29.

Proof of Theorem 1

Case 1: $k \geq c$.

Let $\lambda(\theta)$ be a one point distribution for θ such that $\theta = 0$ WP1.

Let ρ_λ denote the average risk associated with λ and loss function (2). Then

$$\rho_\lambda = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-k\theta^2 + c(\delta-\theta)^2} p(v)dv\, d\lambda(\theta)$$

$$= \int_{-\infty}^{\infty} e^{c\delta^2} p(v)dv.$$

ρ_λ is clearly minimized for $\delta_\lambda(z) = 0$. The associated risk function for the above Bayes estimate is

$$r[\delta_\lambda,\theta] = e^{-\theta^2(k-c)},$$

and is maximized for $\theta = 0$.

Application of Lemma 1 shows $\delta^*_+(z) = 0$ to be a minimax estimate and $\theta = 0$ WP1 to be a least favorable prior distribution for θ.

Case 2: $k < c < k + \frac{1}{2\sigma^2}$.

The conditional risk function associated with $\delta(z) = az$, when a is a constant, is computed to be

$$r^+[\delta,\theta] = \frac{1}{\sqrt{1-2ca^2\sigma^2}} e^{-\theta^2[k - c(a-1)^2 - \frac{2\sigma^2a^2(a-1)^2c^2}{1 - 2ca^2\sigma^2}]}. \tag{5}$$

Now since $k < c$, the minimax estimate over linear rules will be determined by the smallest value of a for which

$$k - c(a-1)^2 - \frac{2\sigma^2a^2(a-1)^2c^2}{1 - 2ca^2\sigma^2} = 0.$$

Routine solution of the above equation gives

$$a^* = \frac{c - \sqrt{kc(1 + 2k\sigma^2 - 2c\sigma^2)}}{2kc\sigma^2 + c}. \tag{6}$$

Simple algebraic manipulations show that

$$0 < a^* < 1, \text{ and}$$

$$1 - 2ca^{*2}\sigma^2 > 0.$$

Since $r^+[\delta^*_+,\theta]$ is independent of θ, $\delta^*_+(z) = a^*z$ is an equalizer rule (constant risk).

We next compute a Bayes rule $\delta_\lambda(z)$ for θ with respect to a prior distribution $\lambda(\theta) = N[0,r^2]$ where $v \in N[0,\sigma^2]$ and the loss function is defined by (2).

We seek a decision rule $\delta(z)$ which minimizes

$$\rho_\lambda = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{c(\delta-\theta)^2 - k\theta^2}\, p(\theta|z)d\theta\; p(z)dz. \tag{7}$$

Now since the inner integral in (7) is strictly positive, we may equivalently minimize

$$\int_{-\infty}^{\infty} e^{c(\delta-\theta)^2 - k\theta^2}\, e^{-\frac{(\theta-[r^2/(r^2+\sigma^2)]z)^2}{2r^2\sigma^2/(r^2+\sigma^2)}}\, d\theta. \tag{8}$$

Rewriting (8) and suppressing multiplicative constants, we may equivalently minimize

$$\int_{-\infty}^{\infty} e^{c(\delta-\theta)^2}\, e^{-\frac{(\theta-\alpha z)^2}{\beta}}\, d\theta\ , \tag{9}$$

where

$$\alpha = \frac{r^2}{2r^2\sigma^2k + r^2 + \sigma^2}\ , \text{ and}$$

$$\beta = \frac{1}{\sqrt{k + \frac{r^2+\sigma^2}{2r^2\sigma^2}}}\ .$$

Expression (9) can be identified as a symmetric convex loss function, $e^{+c(\delta-\theta)^2}$, and a conditional probability density for θ,

$$f(\theta|z) = \frac{1}{\sqrt{\pi\beta}}\, e^{-\frac{(\theta-\alpha z)^2}{\beta}}\ ,$$

which is symmetric about its conditional mean. Hence, Lemma 2 applies and we have the unique (up to equivalence) Bayes rule,

$$\delta_\lambda(z) = \frac{r^2}{2r^2\sigma^2k + r^2 + \sigma^2}\, z. \tag{10}$$

Noting that the above rule is linear, we may equate it to a^*z. We have,

$$a^* = \frac{r^2}{2r^2\sigma^2k + r^2 + \sigma^2}\ .$$

Solving for r^2, we obtain

$$r^2 = \frac{a^*\sigma^2}{1 - a^* - 2\sigma^2ka^*}\ . \tag{11}$$

Inspection of (6) shows that for fixed k,

$$a^* < \frac{1}{2\sigma^2 k + 1}$$

when $k < c < k + 1/2\sigma^2$. Therefore the right-hand side of equation (11) is always positive.

Thus with $\theta \in N[0, r^2]$ and r^2 given by (11), we have a Bayes rule with constant risk and may immediately use Lemma 1 to conclude $\delta^*(z) = a^* z$ is globally minimax.

Admissibility for Cases 1 and 2 follows from the uniqueness (up to equivalence) of the Bayes rules. See Ferguson [5], p. 60.

Case 3: $c = k + \frac{1}{2\sigma^2}$.

Let $\lambda_n(\theta)$ denote a sequence of distributions for θ such that $\theta \in N[0, n^2]$. Let

$$\delta^*_+(z) = \frac{1}{2k\sigma^2 + 1} z. \tag{12}$$

This rule is an equalizer rule with constant conditional risk r_c,

$$r_c = \frac{2k\sigma^2 + 1}{\sqrt{4k^2\sigma^4 + 2k\sigma^2}} \tag{13}$$

We have from equation (10) that the Bayes rule with respect to $\lambda_n(\theta)$ is

$$\delta_{\lambda_n}(z) = \frac{n^2}{2n^2\sigma^2 k + n^2 + \sigma^2} z. \tag{14}$$

Let ρ_n denote the Bayes risk associated with δ_{λ_n}. We compute $\lim_{n\to\infty} \rho_n$:

$$\lim_{n\to\infty} \rho_n = \lim_{n\to\infty} \int_{-\infty}^{\infty} r^+[\delta_{\lambda_n}, \theta] d\lambda(\theta)$$

$$= \frac{2k\sigma^2 + 1}{\sqrt{4k^2\sigma^4 + 2k\sigma^2}}$$

Hence, by Lemma 3, $\delta^*_+(z)$ is minimax.

To prove the admissibility of $\delta^*_+(z)$ we have shown in [6] that for any $\delta(z)$ which is bounded on a finite interval and for which $r[\delta, \theta]$ exists, $r[\delta, \theta]$ is continuous in θ.

Assume $\delta^*_+(z)$ defined by (12) is not admissible. This implies that a $\delta^{*\prime}_+$ exists such that

$$r[\delta^{*\prime}_+, \theta] \leq \frac{2k\sigma^2 + 1}{\sqrt{4k^2\sigma^4 + 2k\sigma^2}}$$

with strict inequality for at least one value of θ. But by the continuity of r, we have

$$r[\delta_{+}^{*'},\theta] \leq \frac{2k\sigma^2+1}{\sqrt{4k^2\sigma^4+2k\sigma^2}} - \mathcal{E} \tag{15}$$

for all θ in some interval $[\theta_1,\theta_2]$ and some $\mathcal{E} > 0$.

Equation (15) can be used to establish the following limit:

$$\lim_{n\to\infty} \frac{r_c^2 - \rho_n'^2}{r_c^2 - \rho_n^2} = \infty , \tag{16}$$

where r_c is defined by (13), ρ_n' is the average risk associated with $\delta_{+}^{*'}$, $\theta \in N[0,n^2]$, and ρ_n is Bayes risk with respect to δ_{λ_n} defined by (14).

Equation (16) implies $\rho_n' < \rho_n$ for sufficiently large n, a contradiction since ρ_n is a Bayes risk. Details of this argument are contained in [6].

Case 4: $c > k + \frac{1}{2\sigma^2}$.

The following argument shows that all rules $\delta(z)$ have unbounded risk.

$$\sqrt{2\pi}\,\sigma\, r[\delta,\theta] = \int_{-\infty}^{\infty} e^{\mathcal{E}\theta^2}\, e^{c\delta^2 - \frac{z^2}{2\sigma^2} + (z/\sigma^2 - 2c\delta)\theta}\, dz, \text{ where } \mathcal{E} = c-k-\frac{1}{2\sigma^2} > 0$$

$$> \int_{0}^{1} e^{\mathcal{E}\theta^2}\, e^{-1/2\sigma^2}\, e^{s\theta}\, dz, \text{ where } s = \inf\{z/\sigma^2 - 2c\delta(z) : z \in [0,1]\}$$

$$= e^{-1/2\sigma^2}\, e^{\mathcal{E}\theta^2 + s\theta} .$$

REFERENCES

1. T. Basar and M. Mintz, "On a Minimax Estimate for the Mean of a Normal Random Vector Under a Generalized Quadratic Loss Function," The Annals of Statistics, Vol. 1, No. 1, pp. 127-134, 1973.

2. D. Jacobson, "Optimal Stochastic Linear Systems with Exponential Performance Criteria and Their Relation to Deterministic Differential Games," IEEE Transactions on Automatic Control, Vol. 18, No. 2, pp. 124-131, 1973.

3. E. L. Lehmann, Notes on the Theory of Estimation, Berkeley, California, University of California Press, 1966.

4. A. J. Viterbi, Principles of Coherent Communication, New York, N. Y., McGraw-Hill, 1966.

5. T. S. Ferguson, Mathematical Statistics - A Decision Theoretic Approach, New York, N. Y., Academic Press, 1967.

6. D. Stelte and M. Mintz, "On Minimax Parameter Estimation Under Exponentiated Generalized Quadratic Loss," submitted for publication.

ACKNOWLEDGMENT

The authors wish to express their gratitude for the following support. This work was supported in part by the following agencies and institutions:

By the U. S. Air Force under Grant AFOSR-68-1579D, and the Joint Services Electronics Program under Contract DAAB-07-72-C-0259 with the Coordinated Science Laboratory, University of Illinois, Urbana, Illinois.

ESTIMATION OF INITIAL CONDITIONS IN LINEAR SYSTEMS*

J. S. MEDITCH
School of Engineering
University of California
Irvine, California 92664

G. H. HOSTETTER
Electrical Engineering Department
California State University
Long Beach, California 90815

ABSTRACT

Techniques for estimating the initial conditions of deterministic and stochastic linear systems are presented and investigated. Difficulties in implementation of existing and proposed algorithms for initial condition estimation are noted and methods for overcoming these difficulties are formulated.

1. INTRODUCTION

In the analysis of system performance, there exist isolated time points at which it is desirable to have a more accurate estimate of the system's state than can be obtained by an observer or a filter. Such situations arise, for example, at a staging point for a spacecraft launch or when a failure occurs in a power system. Another case in point is that where study of a metabolic process requires highly accurate knowledge of the initial concentrations of the reactants immediately following the administration of chemotherapy.

The fundamental problem in such situations is that of estimating the initial conditions of a dynamic system from partial, and generally noisy, measurements of the system's state. The classical solution is well-known and is due to Gauss [1]. More recently, the problem has been examined for stochastic linear systems by Rauch [2] and Meditch [3]. The only known contemporary study of this problem for deterministic linear systems is the one by Hostetter and Meditch [4] and is based on observer theory [5, 6].

In this paper, we examine a number of procedures which can be employed to solve the initial condition estimation problem for both deterministic and stochastic linear systems. Our purpose is to discern the advantages and disadvantages of these procedures. For simplicity and ease of presentation, much of our work will deal with constant-coefficient, continuous-time systems. In such cases, the extension to time-varying and/or discrete-time situations is straightforward, but perhaps tedious. However, where it is convenient and expeditious to do so, we shall treat the time-varying case.

We consider deterministic systems in Section 2 where four different methods of initial condition estimation are studied. The stochastic case we examine in Section 3 paying particular attention to a stability question associated with the fixed-point smoother [3], and the use of the augmentation approach [7] to reduce the initial condition estimation problem from one of smoothing to one of filtering. In Section 4, we present the conclusions of our study.

---

*Research supported by the U.S. Air Force Office of Scientific Research, Air Force Systems Command, under Grant No. AFOSR 71-2116, and by the National Science Foundation under Grant No. GK 39735.

2. DETERMINISTIC CASE

We examine here the problem of estimating the initial state of the system

$$\dot{x} = Ax$$

$$y(t) = Cx(t) \qquad t \geqslant 0 \tag{1}$$

where x is an n-vector, the state; y is an m-vector, the measurement; and (A, C) is completely observable.

Identity Observer with Backward Extrapolation. The identity observer [5, 6] for the system in Eq. (1) is

$$\dot{z} = Az + K(y - Cz) \qquad t \geqslant 0 \tag{2}$$

where K is an n x m matrix whose choice permits arbitrary assignment of the observer eigenvalues.

Defining the observer error to be $\tilde{x}(t|t) = x(t) - z(t)$, we have

$$\dot{\tilde{x}}(t|t) = (A - KC)\,\tilde{x}(t|t) \qquad t \geqslant 0 \tag{3}$$

whose solution is

$$\tilde{x}(t|t) = e^{(A-KC)t}\,\tilde{x}(0|0) \tag{4}$$

A suitable choice of K guarantees the exponential asymptotic stability of the system in Eq. (3) with $\lim\limits_{t\to\infty} \tilde{x}(t|t) = 0$

For any fixed t > 0, a simple way of estimating x(0) is via reverse-time integration of the differential equation in Eq. (1) from t to 0 with z(t) as the boundary condition. Denoting such an estimate as $\hat{x}(0|t)$ to indicate that it is an approximation of x(0) which is based on measurement data over [0, t], we have

$$\hat{x}(0|t) = e^{-At}\,z(t) \tag{5}$$

The solution of Eq. (2) can be written

$$z(t) = e^{At}\,z(0) + \int_0^t e^{A(t-\tau)}K[y(\tau) - Cz(\tau)]\,dt$$

whereupon it follows that

$$\hat{x}(0|t) = z(0) + \int_0^t e^{-A\tau}K[y(\tau) - Cz(\tau)]\,d\tau. \tag{6}$$

Eq. (6) provides a forward-time, sequential algorithm for estimating x(0).

Let us now consider the asymptotic behavior of the estimation error associated with this algorithm. Defining this error as $\tilde{x}(0|t) = x(0) - \hat{x}(0|t)$, we have from Eqs. (6) and (4), and the definition of observer error that

$$\tilde{x}(0|t) = x(0) - z(0) - \int_0^t e^{-A\tau}K[y(\tau) - Cz(\tau)]d\tau$$

$$= \tilde{x}(0|0) - \int_0^t e^{-A^T\tau} KC e^{(A-KC)\tau} d\tau \, \tilde{x}(0|0)$$

$$= \left[I - \int_0^t e^{-A^T\tau} KC e^{(A-KC)\tau} \, d\tau \right] \tilde{x}(0|0) \tag{7}$$

where I is the n x n identity matrix.

A proper choice of K will insure the existence of the integral in Eq. (7) for t = ∞. From Bellman [8, p. 175], we then have that

$$\int_0^\infty e^{-A^T\tau} KC e^{(A-KC)\tau} \, d\tau = I$$

so that $\lim_{t\to\infty} \tilde{x}(0|t) = 0$.

In spite of this result, the use of the algorithm in Eq. (6) is ill-advised in practice. The difficulty stems from the matrix exponential in the integrand in Eq. (6) whose calculation via any numerical technique is unstable if A has any left half-plane eigenvalues. This difficulty is readily traced to the use of reverse-time extrapolation, Eq. (5), from which the algorithm was derived. For indeed, if the system in Eq. (1) is asymptotically stable, reverse-time integration with finite precision arithmetic can lead to serious errors. On the other hand, we note that the above defect is absent if A has only right half-plane eigenvalues. In this case, the stability of the associated observer, Eq. (2), is assured due to the freedom in specifying K, and, thereby, the observer eigenvalues.

Observers with Sampling. The above difficulty can be overcome by the introduction of sampling and the use of a second observer. The approach here was first introduced in [4]; it is summarized below.

Consider the construction of an observer for $\tilde{x}(t|t)$. Defining F = A - KC, we have from Eq. (3) that

$$\dot{\tilde{x}}(t|t) = F\tilde{x}(t|t). \tag{8}$$

Since $y(t) - Cz(t) = C[x(t) - z(t)] = C\tilde{x}(t|t)$ is a signal that is directly available in the identity observer defined by Eq. (2), we have the "new" measurement equation.

$$\zeta(t) = C\tilde{x}(t|t) \tag{9}$$

Further, since observability is invariant under state variable feedback (F, C) is completely observable if (A, C) is. Hence there exists an observer

$$\dot{w} = Pw + M\zeta$$

for the system defined by Eqs. (8) and (9). The corresponding observer error is governed by the differential equation

$$\dot{\tilde{x}} - \dot{w} = (F - MC)\,\tilde{x} - Pw$$

Setting P = F - MC, we have

$$\dot{\tilde{x}} - \dot{w} = (F - MC)\,(\tilde{x} - w) \tag{10}$$

whose eigenvalues can be assigned arbitrarily by fixing M since (F, C) is completely observable.

The solution of Eq. (10) is

$$[x(t) - z(t)] - w(t) = e^{(F-MC)t} \{[x(0) - z(0)] - w(0)\}$$

$$= e^{Pt} \{[x(0) - z(0)] - w(0)\}. \tag{11}$$

Now assume that we set $w(0) = z(0) = 0$. As a consequence, we have first from Eq. (4) that

$$x(t) - z(t) = e^{Ft} x(0) \tag{12}$$

and secondly from Eq. (11) that

$$[x(t) - z(t)] - w(t) = e^{Pt} x(0) \tag{13}$$

Substituting into Eq. (13) from Eq. (12) and rearranging terms, we obtain

$$(e^{Ft} - e^{Pt})\, x(0) = w(t)$$

from which it follows that an estimate of $x(0)$ is given by

$$\hat{x}(0|t) = (e^{Ft} - e^{Pt})^{-1} w(t) \tag{14}$$

for all $t > 0$ for which the indicated inverse exists. One can now introduce sampling at a set of time points $0 < t_1 < t_2 \ldots$ to obtain

$$x(0|t_i) = \left(e^{Ft_i} - e^{Pt_i} \right)^{-1} w(t_i) \tag{15}$$

for $i = 1, 2, \ldots$.

Since the eigenvalues of $F = A - KC$ and $P = F - MC = A - KC - MC$ can be assigned arbitrarily and independently by specifying K and M, respectively, the matrix exponentials in Eq. (15) are well-behaved. The difficulty comes, of course, in the conditioning of the matrix in Eq. (15) and calculation of its inverse.

This latter difficulty can be obviated if the eigenvalues of F are chosen to have large negative real parts relative to the real parts of the eigenvalues of P. In this case, there will exist a time $T > 0$ for which

$$e^{FT} \cong 0 \text{ and } e^{PT} \cong I$$

for which Eq. (15) becomes $\hat{x}(0|T) \cong - w(T)$.

Analog simulations of the latter approach have been conducted with excellent results [4]. The approach does, however, suffer from the fact that it is a "one shot" affair, and does not incorporate additional measurement data as they become available.

Algebraic and Least-Squares Approaches. The most direct way to determine $x(0)$ is algebraically utilizing n independent measurements. For convenience, assume that y is a scalar. Then sampling the output at n distinct time points $0 \leqslant t_1 < t_2 < \ldots t_n$, one obtains from Eqs. (1) and (2) the system

$$\begin{bmatrix} y(t_1) \\ \cdot \\ \cdot \\ \cdot \\ y(t_n) \end{bmatrix} = \begin{bmatrix} Ce^{At_1} \\ \cdot \\ \cdot \\ \cdot \\ Ce^{At_n} \end{bmatrix} x(0) \tag{16}$$

which is invertible to yield x(0) if and only if (A, C) is completely observable.

Theoretically, this method gives x(0) exactly. We can expect it to work well as long as errors in calculating the matrix exponentials can be bounded when A has any right half-plane eigenvalues, and modeling and measurement errors are small. However, the approach does not take advantage of the redundant, but practically useful, measurement data which are available.

The latter can be accomplished by adopting the classical least-squares viewpoint and utilizing well-known matrix inversion lemmas to derive a forward-time, sequential algorithm for estimating x(0).

Again, for convenience, assume that y is a scalar. Also assume that y has been sampled at k + 1 distinct time points $0 \leqslant t_1 < t_2 < \ldots < t_{k+1}$, $k \geqslant n$. The problem is to determine x(0) such that

$$J = \frac{1}{2} \sum_{i=1}^{k+1} \left[y(t_i) - Ce^{At_i} x(0) \right]^2$$

is minimized.

Setting the gradient of J with respect to x(0) equal to zero and employing some tedious, but well-known, matrix manipulations, one arrives at the sequential algorithm

$$\hat{x}(0|t_{k+1}) = \hat{x}(0|t_k) + L(t_{k+1}) \left[y(t_{k+1}) - Ce^{At_{k+1}} \hat{x}(0|t_k) \right] \tag{17}$$

$$L(t_{k+1}) = P(t_k) e^{A't_{k+1}} C' \left[Ce^{At_{k+1}} P(t_k) e^{A't_{k+1}} C' + I \right]^{-1} \tag{18}$$

$$P(t_{k+1}) = \left[I - L(t_{k+1}) Ce^{At_{k+1}} \right] P(t_k) \tag{19}$$

where prime denotes the matrix transpose and

$$P(t_k) = \left[\sum_{i=1}^{k} e^{A't_i} C'Ce^{At_i} \right]^{-1} \tag{20}$$

This algorithm can be initialized at k = n utilizing the results from the algebraic approach defined in Eq. (16). The x(0) obtained there becomes $\hat{x}(0|t_n)$ and $P(t_n)$ is obtained according to Eq. (20) by premultiplying the matrix in Eq. (16) by its transpose and inverting the result.

It is no surprise that the algorithm in Eqs. (17) through (19) has the same structure as the Kalman filter since the latter can be derived, at least formally, in the same way.

Time-Varying Observers. In this fourth approach, we let $\psi(t) = x(0)$ in which case the model in Eq. (1) can be represented as

$$\dot{\psi} = 0$$

$$y(t) = Ce^{At}\ \psi(t) \qquad t \geqslant 0 \tag{21}$$

A time-varying identity observer for the system is specified by

$$\dot{\hat{\psi}} = K(t)\left[y(t) - Ce^{At}\ \hat{\psi}\right] \tag{22}$$

for $t \geqslant 0$ with $\hat{\psi}(0)$ arbitrary and $K(t)$ an $n \times m$ time-varying matrix. Alternately, we have

$$\dot{\hat{\psi}} = -\ K(t)Ce^{At}\ \ \hat{\psi} + K(t)\ y(t) \tag{23}$$

and it is clear that both the Lyapunov and input/output stability of the observer are governed by the choice of $K(t)$.

The corresponding estimation error equation is

$$\dot{\tilde{\psi}} = -\ K(t)Ce^{At}\ \tilde{\psi} \qquad t \geqslant 0 \tag{24}$$

where $\tilde{\psi}(t) \triangleq x(0) - \hat{\psi}(t)$.

While relatively attractive, this approach calls for synthesis of a time-varying linear system to achieve say a prescribed degree of uniform asymptotic stability. The latter is necessary as seen from Eq. (24) to force the estimation error to converge to zero at a suitable rate. At this point, the problem is an open one since it is not known under what conditions on A and C, if any, there exists $K(t)$ which can guarantee a specified degree of stability. The stochastic version of this approach, which is given in the next section, will give some insight into this question.

3. STOCHASTIC CASE

In this section, we shall consider systems which can be described formally by

$$\dot{x} = F(t)x + w(t)$$

$$z(t) = H(t)\ x(t) + v(t) \qquad t \geqslant t_0 \tag{25}$$

where x is an n-vector, the state; z an m-vector, the measurement; w an n-vector, the system disturbance; and v an m-vector, the measurement error. The matrices $F(\cdot)$ and $H(\cdot)$ are assumed to be continuous.

We take $\{w(t),\ t \geqslant t_0\}$ and $\{v(t),\ t \geqslant t_0\}$ to be zero mean, Gaussian, white noises, independent of each other and of the initial state $x(t_0)$, with continuous, positive definite covariance matrices $Q(t)$ and $R(t)$, respectively. We assume that $x(t_0)$ is a zero mean, Gaussian, random vector with positive-definite covariance matrix $P(t_0)$; that $(F(t),\ Q^{\frac{1}{2}}(t))$ is completely controllable; and that $(F(t),\ H(t))$ is completely observable.

These conditions guarantee the stability of the Kalman filter:

$$\dot{\hat{x}}(t|t) = F(t)\ \hat{x}(t|t) + K(t)\ [z(t) - H(t)\ \hat{x}(t|t)] \tag{26}$$

$$K(t) = P(t|t)\ H'(t)\ R^{-1}(t) \tag{27}$$

$$\dot{P}(t|t) = F(t)P(t|t) + P(t|t)F'(t) - P(t|t)H'(t)R^{-1}(t)H(t)P(t|t) + Q(t) \tag{28}$$

where $t \geqslant t_0$, $\hat{x}(t_0|t_0) = 0$, $P(t_0|t_0) = P(t_0)$, and $P(t|t)$ is the covariance matrix of the filtering error $\tilde{x}(t|t) \triangleq x(t) - \hat{x}(t|t)$.

Fixed-Point Smoother. A well-known algorithm [3] for sequential computation of an estimate of $x(t_0)$ is defined by the relation

$$\dot{\hat{x}}(t_0|t) = B(t)\ K(t)\ [z(t) - H(t)\ \hat{x}(t|t)] \tag{29}$$

where $t \geqslant t_0$ and the n x n matrix B is the solution of

$$\dot{B} = -B\ [F(t) + Q(t)\ P^{-1}(t|t)] \tag{30}$$

with $B(t_0) = I$.

The solution of Eq. (29) is obviously

$$\hat{x}(t_0|t) = \int_{t_0}^{t} B(\tau)\ K(\tau)\ [z(\tau) - H(\tau)\ \hat{x}(\tau|\tau)]d\tau. \tag{31}$$

We note that Eq. (6) is the deterministic version of Eq. (30).

Two problems arise here. First, the determination of B requires knowledge of $P^{-1}(t|t)$ which, in turn, requires implementation of the differential equation [9, p. 267]:

$$\dot{P}^{-1}(t|t) = - P^{-1}(t|t)\ F(t) - F'(t)\ P^{-1}(t|t) + H'(t)\ R^{-1}(t)\ H(t) - P^{-1}(t|t)\ Q(t)\ P^{-1}(t|t)$$

for $t \geqslant t_0$ with $P^{-1}(t_0|t_0) = P^{-1}(t_0)$.

The second problem has to do with the behavior of the matrix B(t). Parenthetically, we remark that the signal $K(t)\ [z(t) - H(t)\ \hat{x}(t|t)]$ is directly available from the Kalman filter and poses no difficulty.

We now show that, subject to the above assumptions, the calculation of B(t) possesses the same stability properties as the Kalman filter. To do this, we introduce the transformation $C(t) = P(t|t)\ B'(t)$ and obtain with the aid of Eqs. (28) and (30) the result

$$\dot{C}(t) = [F(t) - K(t)\ H(t)]\ C(t)$$

where $t \geqslant t_0$ and $C(t_0) = P(t_0|t_0) = P(t_0)$. The assertion follows trivially.

However, if there is no system disturbance, $Q(t) = 0$ and Eq. (3) becomes

$$\dot{B} = - BF(t).$$

In this case, B(t) inherits the stability properties of -F(t) and we have the same situation as in Eq. (6) for the deterministic case. Here, however, we have a way out!

Let $\Phi(t,t_0)$ denote the state transition matrix for the disturbance-free system $\dot{x} = F(t)x$, $t \geqslant t_0$. Further, let $\psi(t) = x(t_0)$ whereupon the system model becomes

$$\dot{\psi} = 0$$

$$z(t) = H(t)\ \Phi(t, t_0)\ \psi(t) + v(t)$$

with $t \geqslant t_0$. The Kalman theory is directly applicable to this problem. Noting that $\hat{\psi}(t|t) = \hat{x}(t_0|t)$, we have

$$\dot{\hat{x}}(t_0|t) = K(t)\ [z(t) - H(t)\ \Phi(t, t_0)\ \hat{x}(t_0|t)] \tag{32}$$

$$K(t) = P(t|t)\ [H(t)\ \Phi(t,t_0)]'R^{-1}(t) \tag{33}$$

$$\dot{P}(t|t) = -\ P(t|t)\ \Phi'(t,t_0)\ H'(t)\ R^{-1}(t)\ H(t)\ \Phi(t,t_0)\ P(t|t) \tag{34}$$

where $t \geqslant t_0$, $\hat{x}(t_0|t_0) = 0$ and $P(t_0|t_0) = P(t_0)$.

The result in Eq. (32) is analogous to the time-varying observer in Eq. (22). In contrast to the latter, however, the synthesis problem here is completely resolved by Eqs. (33) and (34), and, moreover, so is the stability issue.

One could use the above results as a guide in tackling the deterministic problem. For example, equal weighting could be assigned to the measurements by fixing $R(t) = I$ and $P(t_0)$ could be made a diagonal matrix with each element equal to say 1000 to associate a large uncertainty with the *a priori* knowledge of $x(0)$. However, the ramifications of such an approach remain to be examined.

Augmentation. In [7], Biswas and Mahalanabis examined the fixed-point smoothing problem as one of higher-dimensional Kalman filtering. The approach avoids the calculation of $P^{-1}(t|t)$ and $B(t)$ which are required in the algorithm of Eqs. (29) and (30), but does require solution of a 2n-dimensional filtering problem. It is, nevertheless, computationally more attractive than the previous algorithm.

To begin, we augment the state of the system of Eq. (25) with $\psi(t) = x(t_0)$ defining the new state to be

$$y = \begin{bmatrix} x \\ \psi \end{bmatrix}$$

The system equations then become

$$\dot{y} = A(t)y + \eta(t)$$

$$z(t) = C(t)y(t) + v(t)$$

for $t \geqslant t_0$ where

$$A(t) = \begin{bmatrix} F(t) & 0 \\ 0 & 0 \end{bmatrix},\quad C(t) = [H(t) \quad 0] \quad \text{and} \quad \eta(t) = \begin{bmatrix} w(t) \\ 0 \end{bmatrix}$$

and the 0's are null matrices and vectors of suitable order.

The covariance matrix of $\eta(t)$ is

$$\overline{Q}(t) = \begin{bmatrix} Q(t) & 0 \\ 0 & 0 \end{bmatrix}$$

and that of $y(t_0)$ is

$$P(t_0) = \begin{bmatrix} \Pi(t_0) & \Pi(t_0) \\ \Pi(t_0) & \Pi(t_0) \end{bmatrix}$$

where $\Pi(t_0)$ is the covariance matrix of $x(t_0)$. All other quantities are as before.

Kalman filter theory can now be applied directly to this problem, it again being clear that $\hat{\psi}(t|t) = \hat{x}(t_0|t)$.

The distinct advantage of this approach over the others for the stochastic problem becomes apparent when we express the filtering problem solution in component form. To do this, we let

$$P(t) = \begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{21}(t) & P_{22}(t) \end{bmatrix}$$

denote the covariance matrix of the estimation error $\tilde{y}(t|t) = y(t) - \hat{y}(t|t)$. In this form, $P_{11}(t)$ is the covariance matrix of $\tilde{x}(t|t) = x(t) - \hat{x}(t|t)$, $P_{22}(t)$ the covariance matrix of $\tilde{x}(t_0|t) = x(t_0) - \hat{x}(t_0|t)$, and $P_{12}(t)$ the cross-covariance matrix of $\tilde{x}(t|t)$ and $\tilde{x}(t_0|t)$. We thus have

$$\dot{P}_{11} = F(t)P_{11} + P_{11}F'(t) - P_{11}H'(t)R^{-1}(t)H(t)P_{11} + Q(t) \tag{35}$$

$$\dot{P}_{12} = F(t)P_{12} - P_{11}H'(t)R^{-1}(t)\, H(t)\, P_{12} \tag{36}$$

and

$$\dot{P}_{22} = -P'_{12}\, H'(t)\, R^{-1}(t)\, H(t)\, P_{12} \tag{37}$$

for $t \geqslant t_0$ subject to $P_{11}(t_0) = P_{12}(t_0) = P_{22}(t_0) = \Pi(t_0)$.

Clearly, Eq. (35) is independent of Eqs. (36) and (37), and would have to be implemented in any case to realize the Kalman filter for $\hat{x}(t|t)$. The latter two equations are then needed for the fixed-point smoothing solution. The relative simplicity in component form arises from the sparsity of the matrices A(t), C(t) and $\overline{Q}(t)$.

The corresponding filter equations are easily shown to be

$$\dot{\hat{x}}(t|t) = F(t)\,\hat{x}(t|t) + K_1(t)\,[z(t) - H(t)\,\hat{x}(t|t)]$$

and

$$\dot{\hat{x}}(t_0|t) = K_2(t)\,[z(t) - H(t)\,\hat{x}(t|t)]$$

where

$$K_1(t) = P_{11}(t)\, H'(t)\, R^{-1}(t)$$

and

$$K_2(t) = P'_{12}(t)\, H'(t)\, R^{-1}(t)$$

for $t \geqslant t_0$ where $\hat{x}(t_0|t_0) = 0$ in both equations.

If one desires to estimate only $\hat{x}(t_0|t)$, then the algorithm of Eqs. (32) through (34) is clearly preferable to the present one if there is no system disturbance since only an n-dimensional filter is needed. However, in practice, one is generally interested in estimates of the current state as well, in which case the augmentation approach of [7], as outlined here, appears to be the most advantageous one presently available.

4. CONCLUSIONS

We have presented four techniques for estimating the initial state of deterministic linear systems and three techniques for stochastic linear systems. For the latter case, the development clearly delineates the practical advantages of the augmentation approach over the other two.

In the former case, the picture is not quite as clear except for the formulation involving the identity observer with backward extrapolation, Eqs. (5) and (6), which must be discounted on the basis of computational instability. The "observers with sampling" approach, while limited to a "single shot", is easy to implement, and is useful if model and measurement errors are negligible. The algebraic and least-squares methods are classical and offer the best general attack on the problem. However, the recursive formulation of Eqs. (17) through (19) is somewhat awkward in its requirement of calculation of the matrix exponential, or equivalently the state transition matrix in the time-varying case. Finally, the time-varying observer of Eq. (22) is attractive for sequential estimation of $x(0)$, but requires care in the selection of $K(t)$ as noted.

REFERENCES

1. K. F. Gauss, "Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections," translated by C. H. Davis, Dover, N.Y., 1963.

2. H. E. Rauch, "Solutions to the Linear Smoothing Problem," IEEE Trans. Auto. Cont., vol. AC-8, October 1963, pp. 371-372.

3. J. S. Meditch, "Optimal Fixed-Point Continuous Linear Smoothing," Proc. 1967 Joint Auto. Cont. Conf., Philadelphia, Pa., June 1967, pp. 249-256.

4. G. H. Hostetter and J. S. Meditch, "Observers for the Computation of Initial Conditions," Proc. Fourth Symp. on Nonlinear Estimation Theory and Its Appl., San Diego, Calif., September 1973.

5. D. G. Luenberger, "Observing the State of a Linear System," IEEE Trans. Mil. Elect., vol. MIL-8, April 1964, pp. 74-80.

6. D. G. Luenberger, "Observers for Multivariable Systems," IEEE Trans. Auto. Cont., vol. AC-11, April 1966, pp. 190-197.

7. K. K. Biswas and A. K. Mahalanabis, "An Approach to Fixed-Point Smoothing Problems," IEEE Trans. Aero. and Elect. Sys., vol. AES-8, September 1972, pp. 676-682.

8. R. Bellman, "Introduction to Matrix Analysis," McGraw-Hill, New York, 1960.

9. J. S. Meditch, "Stochastic Optimal Linear Estimation and Control," McGraw-Hill, New York, 1969.

ON THE SOLUTION OF A MINIMAX TERMINAL STATE ESTIMATION PROBLEM--A HILBERT SPACE APPROACH

GYORGY HEXNER and MAX MINTZ
Coordinated Science Laboratory and
Department of Electrical Engineering
University of Illinois
Urbana, Illinois 61801

ABSTRACT

The solution to a continuous-time LQG minimax terminal state estimation problem is obtained using a Hilbert space approach. The minimax decision rule is determined via the method of least favorable prior distributions. The minimax estimator is shown to be a linear functional of the observed data, and the least favorable prior distribution is shown to be a degenerate Gaussian stochastic process.

1. INTRODUCTION

In a recent paper [1] Basar and Mintz have considered a discrete-time minimax terminal state estimation problem of the LQG variety. The details of the solution described in [1] depend fundamentally on the nature of the topology of E^n -- a topology which is compatible with the discrete-time problem formulation. Because of the specific dependence on the topology of E^n, the minimax theorems developed in [1] cannot be directly extended to include the continuous-time problem formulation. The purpose of the present study is to provide a solution to the continuous-time version of the minimax terminal state estimation problem considered in [1], and to identify the essential structure of the solution which is independent of the nature of the time domain.

Minimax state estimation problems can be regarded as two person zero sum differential games with imperfect state information. Differential games with imperfect state information have recently been studied by various authors, see for example [2,3,4,5,6,7].

The motivation for the terminal state estimation problem considered here is provided by a radar data processing problem. The latter problem is associated with tracking an uncooperative target which is under the control of an intelligent adversary. The only information available to the tracker is the dynamics of the target, the statistical description of the initial state of the target, and the statistics of the additive observation noise corrupting the radar data.

We shall take a minimax approach and assume that the evader, who controls the target, is faced with a compromise between a gain due to the tracker's estimation error and a loss which the evader incurs by using energy to maneuver evasively.

We further assume that the evader makes no measurements of his own state at any time. For example, this is a valid assumption in the situation where one is tracking a blind incoming missile which has been preprogrammed to maneuver evasively. Consequently, the evader's permissible control policies are forced to be open-loop control policies that do not depend functionally on the initial state of the target.

2. PROBLEM STATEMENT

For simplicity, we restrict our discussion here to a scalar problem.

Player I, (the evader) chooses a function $u(t) \in L_2[0,T]$ to control the dynamic system

$$\dot{x}(t) = f(t)x(t) + g(t)u(t) \tag{2.1}$$

where, we assume that $f(t)$ and $g(t)$ are continuous functions defined on $[0,T]$. The initial state, $x(0)$ of the system (2.1) is a zero mean Gaussian random variable with variance q^2.

Player II, (the tracker) observes $z(t)$, where

$$z(t) = h(t)x(t) + v(t). \tag{2.2a}$$

We make the following assumptions about the functions appearing in equation (2.2a):

(i) $h(t)$ is continuous on $[0,T]$;

(ii) there exist positive numbers δ and k such that

$$|h(t)| \geq k \quad \text{when } 0 < T-\delta \leq t \leq T; \tag{2.2b}$$

(iii) $v(t)$ is zero mean Gaussian white noise with covariance function

$$E[v(t)v(\tau)] = r(t)\delta(t-\tau); \tag{2.2c}$$

(iv) $r(t)$ is a positive continuous function on $[0,T]$.

The tracker may choose an arbitrary functional $\Psi(z)$, which maps the observed data set $\{z(t) : t \in [0,T]\}$ into R, to be his estimate for $x(T)$. We assume that the control u is functionally independent of the random variables $\{x(0)$ and $v(t)$, $t \in [0,T]\}$.

Let $L(\Psi,u)$ be the loss function

$$L(\Psi,u) = [x(T) - \Psi]^2 - \int_0^T d(t)u^2(t)dt \tag{2.3}$$

where $d(t)$ is a continuous positive function, defined on $[0,T]$.

We define the conditional risk $R(\Psi,u)$ by

$$R(\Psi,u) = E[L(\Psi,u)|u]. \tag{2.4}$$

The remainder of this paper is concerned with finding an estimate $\Psi_o(z)$ which satisfies

$$\sup_{u \in L_2[0,T]} R(\Psi_o,u) = \inf_{\Psi} \sup_{u \in L_2[0,T]} R(\Psi,u). \tag{2.5}$$

The estimate defined by equation (2.5) is called a minimax estimate.

3. MAIN RESULTS

We rewrite the equations which describe the dynamical system equation (2.1) in integral form:

$$x_1(t) = \Phi(t)\int_0^t \Phi^{-1}(s)g(s)u(s)ds \tag{3.1a}$$

$$x_2(t) = \Phi(t)x(0) \tag{3.1b}$$

$$x(t) = x_1(t) + x_2(t) \tag{3.1c}$$

where, $\Phi(t)$ satisfies the equation

$$\dot{\Phi}(t) = f(t)\Phi(t) \tag{3.1d}$$

$$\Phi(0) = 1.$$

Equation (3.1a) may be further shortened to:

$$x_1(t) = \int_0^T \pi(t,s)u(s)ds \tag{3.2a}$$

where

$$\pi(t,s) = \Phi(t)\Phi^{-1}(s)g(s) \qquad t \geq s \tag{3.2b}$$

$$= 0 \qquad t < s. \tag{3.2c}$$

Theorem 3.1: Consider the Weiner-Hopf equation

$$h(t)R_{xx}(T,t) - \int_0^T a(s)R_{xx}(s,t)h(s)h(t)ds - r(t)a(t) = 0, \tag{3.3a}$$

and define $\beta(t)$ by the equation

$$\beta(t) = \frac{1}{d(t)}\left[\pi(T,t) - \int_0^T h(s)a(s)\pi(s,t)ds\right]. \tag{3.3b}$$

Then either there is a solution of equation (3.3) subject to

$$\int_0^T d(t)\beta^2(t)dt < 1 \tag{3.4a}$$

where $R_{xx}(t,s)$ is the autocorrelation function of the process

$$x(t) = \Phi(t)x(0) \quad ; \; x(0) \in N(0,q^2), \tag{3.4b}$$

or there is a number λ, such that the solution of equation (3.3) satisfies

$$\int_0^T d(t)\beta^2(t)dt = 1, \tag{3.5a}$$

where $R_{xx}(t,s)$ is the autocorrelation function of the process

$$x(t) = \int_0^T \pi(t,s)u(s)ds + \Phi(t)x(0), \tag{3.5b}$$

where $u(t)$ is the stochastic process

$$u(t) = s\lambda\beta(t), \tag{3.5c}$$

$$s \in N(0,1), \tag{3.5d}$$

and

$$x(0) \in N(0,q^2). \tag{3.5e}$$

Theorem 3.2: If $\Psi_o(z)$ is defined by

$$\Psi_o(z) = \int_0^T a(t)z(t)dt \tag{3.6}$$

where $a(t)$ satisfies equations (3.3) and (3.4) or equations (3.3) and (3.5), then $\Psi_o(z)$ is a minimax estimate for $x(T)$.

4. A LEAST FAVORABLE PROBABILITY MEASURE FOR u(t)

In this section, we prove Theorem 3.2.

Definition 4.1: A probability measure μ_o over the space of evader control functions is said to be least favorable if

$$\inf_{\Psi} E_{\mu_o}[R(\Psi,u)] = \sup_{\mu} \inf_{\Psi} E_{\mu}[R(\Psi,u)]. \tag{4.1}$$

We use the next lemma to connect the following concepts: least favorable probability measure, Bayes decision rule, and the value of the game.

Lemma 4.1: If $\Psi_o(z)$ is a Bayes decision rule with respect to μ_o and, for all $u(t) \in L_2[0,T]$

$$R(\Psi_o,u) \leq E_{\mu_o}[R(\Psi_o,u)] \tag{4.2}$$

then the game has a value, $\Psi_o(z)$ is a minimax decision rule, and μ_o is least favorable. The proof of this lemma can be found in ref. [8].

We begin the proof of Theorem 3.2 by restricting our attention to linear decision rules.

$$\Psi_L(z) = \int_0^T a(t)z(t)dt. \tag{4.3}$$

Using $\beta(t)$ defined in equation (3.3b), we rewrite $R(\Psi_L,u)$ as:

$$R(\Psi_L,u) = (\beta,u)^2 - \|u\|^2 + \int_0^T r(t)a^2(t)dt + q^2[\Phi(T) - \int_0^T h(t)a(t)\Phi(t)dt]^2 \tag{4.4}$$

where (β,u) and $\|u\|^2$ denote the inner product and norm with respect to the weighting function $d(t)$.

We consider the two cases referred to in the previous section.

Case 1: $(\|\beta\| < 1)$

Assume that a(t) satisfies equations (3.3) and (3.4) i.e. $\|\beta\| < 1$. We maximize $R(\Psi_L,u)$ over $u \in L_2[0,T]$ as follows:

(Step 1) Assume $\|u\|^2$ satisfies

$$\|u\|^2 = k^2\|\beta\|^2 \tag{4.5a}$$

for some k^2, then, $R(\Psi_L,u)$ is maximized by

$$u(t) = \pm k\beta(t). \tag{4.5b}$$

(Step 2) We maximize $R(\Psi_L, \pm k\beta)$ over k,

$$R(\Psi_L, \pm k\beta) = k^2\|\beta\|^2[\|\beta\|^2-1] + \int_o^T r(t)a^2(t)dt + q^2[\Phi(T)-\int_o^T h(t)a(t)\Phi(t)dt]^2 \tag{4.5c}$$

and find that the maximum is attained at $k=0$, i.e. $R(\Psi_L,u)$ is maximized by $u=0$ a.e. Therefore, we have established the inequality

$$R(\Psi_L,u) \le R(\Psi_L,\theta) = E_{\mu_o}[R(\Psi_L,u)] \tag{4.6}$$

where μ_o is the probability measure which assigns probability one to the zero function θ in $L_2[0,T]$, i.e., w.p.1,

$$u = \theta. \tag{4.7}$$

Since a(t) satisfies the Weiner-Hopf equation (3.3a), Ψ_L defined in equation (4.3) is a linear Bayes decision rule. On the other hand, this is a Linear Quadratic Gaussian (LQG) problem, so that Ψ_L is the Bayes decision rule over the space of all linear and nonlinear functionals. Therefore we have verified the hypothesis of Lemma 4.1, and hence Ψ_o, defined in equation (3.7) is a minimax estimate for x(T).

Case 2: $(\|\beta\| = 1)$

Assume that a(t) satisfies equations (3.3) and (3.5). Following our previous reasoning, we find that $R(\Psi_L,u)$ is maximized by

$$u(t) = k\beta(t) \tag{4.8a}$$

for any k. Now suppose that

$$k = \lambda \tag{4.8b}$$

where λ is the constant in equation (3.5c). Further suppose that u(t) is the stochastic process defined by equation (3.5c), so that

$$E_{\mu_o}[R(\Psi_L,u)] = R(\Psi_L,\lambda\beta) \tag{4.9}$$

where μ_o is the probability measure induced by the stochastic process u(t). Therefore, we have the inequality:

$$R(\Psi_L,u) \leq E_{\mu_o}[R(\Psi_L,u)] = R(\Psi_L,\lambda\beta). \tag{4.10}$$

Since a(t) satisfies the Weiner-Hopf equation (3.3a), Ψ_L is a linear Bayes estimate. On the other hand, this is an LQG problem, so that Ψ_L is the Bayes estimate over the space of all linear and nonlinear functionals. Therefore, we have verified the hypothesis of Lemma 4.1 so that Ψ_o defined in equation (3.6) is a minimax estimate for x(T).

5. DISCUSSION

The solution of the minimax terminal state estimation problem stated in Section 2 has the following salient characteristics:

(i) The tracker's minimax decision rule is a linear functional of the observed data.

(ii) The evader's maximin policy is defined by a degenerate Gaussian stochastic process i.e., a Gaussian process with mulitplicity one.

These results correspond to the general vector solution obtained in [1], where (i) the minimax decision rule for the tracker was linear, and (ii) the maximin policy for the evader was a discrete-time vector valued Gaussian stochastic process.

We recall that the first alternative of Theorem 3.1 implies that $\lambda = 0$. If $\lambda = 0$, then the maximin policy for the evader is to apply zero control, and consequently the related minimax estimate $\Psi_o(z)$ can be realized as a Kalman filter. The multidimensional generalization of this problem has been the subject of a previous paper by one of the authors [9]. In the present paper, inequality (3.4a) takes the place of the condition derived in [9], namely that a certain matrix Riccati equation possesses a bounded solution on the time interval [0,T]. It can also be shown that another Kalman filter can be used to implement the minimax decision rule when the second alternative of Theorem 3.1 holds.

We now show that we have obtained the Nash solution of a particular non-zero sum stochastic differential game. Define the following payoff functions:

$$L_T(\Psi,u) = [x(T) - \Psi(z)]^2 \tag{5.1a}$$

$$L_E(\Psi,u) = [x(T) - \Psi(z)]^2 - \int_o^T d(t)u^2(t)dt, \tag{5.1b}$$

where (5.1a) denotes the tracker's payoff function and (5.1b) denotes the evader's payoff function. Under the Nash solution concept, the objectives of the tracker and the evader are to seek respectively strategies Ψ_o, and μ_o such that for all pairs of decision rules and probability measures (Ψ,μ),

$$E_{\mu_o}[E[L_T(\Psi_o,u)|u]] \leq E_{\mu_o}[E[L_T(\Psi,u)|u]]; \tag{5.2a}$$

$$E_{\mu_o}[E[L_E(\Psi_o,u)|u]] \geq E_{\mu}[E[L_E(\Psi_o,u)|u]]. \tag{5.2b}$$

It follows by inspection that the pair (Ψ_o, μ_o) defined in Sections 3 and 4 also satisfies inequalities (5.2a) and (5.2b). Hence the solution of our original minimax estimation problem is also the solution for this particular non-zero sum stochastic differential game.

We conclude this discussion with the observation that the Hilbert Space approach which was used in the development of the solution to this continuous-time minimax estimation problem also applies to the related discrete-time problem. The Hilbert Space approach used in the present paper is simpler than the approach used in [1] and almost entirely eliminates topological considerations.

APPENDIX A1

We now prove Theorem 3.1. We shall show that λ^2 which is defined in the second part of Theorem 3.1 is the Lagrange multiplier associated with a particular constrained minimization problem. To simplify the ensuing algebra, we use standard Hilbert-space notation. We adhere to the following conventions: symbols which previously represented time functions defined on $[0,T]$, with the time dependence suppressed denote either (i) elements of $L_2[0,T]$, or (ii) operators which map elements of $L_2[0,T]$ into itself by pointwise multiplication. Further, let

$$\alpha(t) = \pi(T,t) \tag{A1.1}$$

and, S be the operator defined by

$$Su \underline{\Delta} \int_0^T \pi(t,s)u(s)ds, \tag{A1.2}$$

$$u \in L_2[0,T].$$

Consider the following minimization problem: minimize the functional $f(a)$, $a \in L_2[0,T]$

$$f(a) \underline{\Delta} (a, d^{-1}ra) + q^2[\Phi(T) - (\Phi, d^{-1}ha)]^2 \tag{A1.3a}$$

subject to the constraint,

$$(d^{-1}\alpha - S^+d^{-1}ha, d^{-1}\alpha - S^+d^{-1}ha) - 1 \leq 0 \tag{A1.3b}$$

where the inner products are taken with respect to the weighting function d(t). We note that

$$d^{-1}\alpha - S^+d^{-1}ha = \beta \tag{A1.4}$$

where $\beta(t)$ was defined in equation (3.3b).

We now show that there exists an $a_o \in L_2[0,T]$ which achieves the minimum of the functional $f(a)$ subject to the constraint (A1.3b).

First, we exhibit an $a_1 \in L_2[0,T]$ which satisfies the inequality

$$(d^{-1}\alpha - S^+d^{-1}ha_1, d^{-1}\alpha - S^+d^{-1}ha_1) \leq \frac{1}{2} . \tag{A1.5}$$

Let M, k_1, k_2 be positive numbers such that

$$|\Phi^{-1}(t)g(t)| \leq M \qquad 0 \leq t \leq T \tag{A1.6a}$$

$$|\Phi(T)| \qquad \leq M \tag{A1.6b}$$

$$k_1 \leq r(t) \leq k_2 \qquad 0 \leq t \leq T \tag{A1.6c}$$

$$k_1 \leq d(t) \leq k_2 \qquad 0 \leq t \leq T. \tag{A1.6d}$$

Since $\Phi(t)$ is a continuous function, there is a number $\delta_1 > 0$ which satisfies,

$$|\Phi(T) - \Phi(t)| \leq \sqrt{\frac{k_1}{4}}\ \frac{1}{T^{\frac{1}{2}}M} \qquad 0 < T-\delta_1 \leq t \leq T. \tag{A1.6e}$$

Let

$$\delta_o = \min(\delta_1, \frac{k_1}{16M^4}, \delta) \tag{A1.7}$$

where we recall that δ was defined in equation (2.2b). Now define $a_1(t)$ as:

$$\begin{aligned} a_1(t) &= \frac{1}{\delta_o h(t)} \qquad T-\delta_o \leq t \leq T \\ &= 0 \qquad\qquad 0 \leq t < T-\delta_o. \end{aligned} \tag{A1.8}$$

Then,

$$\beta(t) = \frac{1}{d(t)}\ \frac{1}{\delta_o}\ [\int_{T-\delta_o}^{T} \Phi(T)-\Phi(s)ds]\Phi^{-1}(t)g(t) \qquad 0 \leq t \leq T-\delta_o \tag{A1.9a}$$

so that

$$|\beta(t)| \leq \frac{1}{d(t)} \sqrt{\frac{k_1}{4}}\ \frac{1}{T^{\frac{1}{2}}M}\ M = \frac{1}{d(t)} \sqrt{\frac{k_1}{4}}\ \frac{1}{T^{\frac{1}{2}}} \qquad 0 \leq t \leq T-\delta_o. \tag{A1.9b}$$

Also,

$$\beta(t) = \frac{1}{d(t)}\ [\Phi(T)\Phi^{-1}(t)g(t) - \frac{1}{\delta_o}\int_{t}^{T} \Phi(s)ds\ \Phi^{-1}(t)g(t)],\ T-\delta_o < t \leq T, \tag{A1.10a}$$

hence,

$$|\beta(t)| \leq \frac{1}{d(t)}\ 2M^2.$$

Therefore, we obtain the inequality for $||\beta||^2$:

$$\begin{aligned} ||\beta||^2 = \int_0^T d(t)\beta^2(t)dt &\leq \int_0^{T-\delta_o} \frac{1}{d(t)}\ \frac{k_1}{4}\ \frac{1}{T}\ dt + \int_{T-\delta_o}^{T} \frac{1}{d(t)}\ 4M^4dt \leq \frac{1}{4} \\ &+ \frac{1}{k_1}\ 4M^4 \cdot \frac{k_1}{16M^4} = \frac{1}{2}. \end{aligned} \tag{A1.11a}$$

Recalling equation (A1.4), we note that

$$(d^{-1}\alpha - S^{+}d^{-1}ha_1,\ d^{-1}\alpha - S^{+}d^{-1}ha_1) = \|\beta\|^2 \le \frac{1}{2} \qquad \text{(A1.11b)}$$

which is inequality (A1.5).

We next observe that:

$$f(a_1) \le P = \frac{k_2}{\delta_o k^2} + 4q^2M^2, \qquad \text{(A1.12)}$$

and therefore, we may place the additional constraint

$$(a,d^{-1}ra) \le P \qquad \text{(A1.13)}$$

on the problem of minimizing f(a) subject to inequality (A1.3b), without altering the minimization problem. The set of a's which satisfy inequality (A1.13) is a weakly compact set. The set defined by inequality (A1.3b) is the inverse image of a closed set by a weakly lower semi-continuous function; hence, it is a weakly closed set. Therefore, the intersection of the sets defined by inequalities (A1.3b) and (A1.13) is a weakly compact set. The functional f(a) is a weakly lower semi-continuous functional; therefore, there exists an $a_o \in L_2[0,T]$ which achieves the minimum of f(a) subject to constraint (A1.3b).

We use the following Lagrange multiplier theorem to establish the existence of λ^2.

Theorem A1.1: (Luenberger [10], p. 217). Let A be a linear vector space, Ω a convex subset of A. Let f be a real-valued convex functional on Ω and G a convex mapping from Ω to R. Assume the existence of a point $a_1 \in \Omega$ for which $G(a_1) < 0$. Let

$$\mu_o = \inf f(a) \text{ subject to } a \in \Omega,\ \ G(a) < 0 \qquad \text{(A1.14a)}$$

and assume μ_o is finite. Then there is a number $z_o \ge 0$ such that

$$\mu_o = \inf_{a\in\Omega} \{f(a) + z_o G(a)\}. \qquad \text{(A1.14b)}$$

Furthermore, if the infinum is achieved in (A1.14a), by an $a_o \in \Omega$ such that $G(a_o) \le 0$, it is achieved by a_o in (A1.14b) and

$$z_o G(a_o) = 0. \qquad \text{(A1.14c)}$$

In our problem, we identify A and Ω with $L_2[0,T]$ and G(a) with the functional on the left hand side of inequality (A1.3b).

Since $z_o \ge 0$, we use instead λ_o^2. Then the functional on the right hand side of equation (A1.14b) becomes

$$(a,d^{-1}ra) + q^2[\Phi(T) - (\Phi,d^{-1}ha)]^2$$
$$+ \lambda_o^2[(d^{-1}\alpha - S^{+}d^{-1}ha,d^{-1}\alpha - S^{+}d^{-1}ha) - 1] \qquad \text{(A1.15)}$$

Since (A1.15) is a positive definite quadratic functional, a necessary and sufficient condition for a_o to minimize (A1.15) is that the first variation about a_o vanish. This condition leads to the equation:

$$ra_o - q^2\Phi(T)h\Phi + q^2(\Phi,d^{-1}ha_o)h\Phi - \lambda_o^2 hS[d^{-1}\alpha - S^+d^{-1}ha_o] = \theta. \qquad \text{(A1.16)}$$

We next show that the pair a_o, λ_o satisfies Theorem 3.1. We consider two cases: (i) a_o is an interior point and (ii) a_o is a boundary point of the set defined by inequality (A1.3b). These two cases correspond to the two alternatives in Theorem 3.1; that is, either equations (3.3) and (3.4) or equations (3.3) and (3.5) have a solution.

First, suppose that a_o is an interior point. Then equation (A1.14c) implies that

$$\lambda_o = 0. \qquad \text{(A1.17)}$$

For the first alternative of Theorem 3.1, the Weiner-Hopf equation (3.3a) can be rewritten as:

$$-q^2\Phi(T)h\Phi + q^2(\Phi,d^{-1}ha)h\Phi + ra = \theta \qquad \text{(A1.18)}$$

which is the same as equation (A1.16) with $\lambda_o = 0$), so that if a_o is an interior point, a_o satisfies equation (3.3) and equation (3.4) (recall equation (A1.4)).

Now suppose that a_o is a boundary point. For the second alternative of Theorem 3.1, using equations (3.5b) and (A1.4), the Weiner-Hopf equation (3.3a) can be rewritten as:

$$\begin{aligned} &ra - q^2\Phi(T)h\Phi + q^2(\Phi,d^{-1}ha)h\Phi \\ &\quad -\lambda^2(d^{-1}\alpha - S^+d^{-1}ha, d^{-1}\alpha - S^+d^{-1}ha)hS[d^{-1}\alpha - S^+ha] = \theta. \qquad \text{(A1.19)} \end{aligned}$$

Now, substituting a_o for a and λ_o for λ, and recalling that for a_o, a boundary point, inequality (A1.3b) becomes an equality, (and hence equation (3.5a) is satisfied). Thus equation (A1.19) reduces to equation (A1.16) so that the pair a_o and λ_o satisfies equations (3.3) and (3.5).

This completes the proof of Theorem 3.1.

REFERENCES

1. T. Basar and M. Mintz, "Minimax Terminal State Estimation for Linear Plants with Unknown Forcing Functions," Int. J. Control, Vol. 16, pp. 49-70, Aug. 1972.

2. R. D. Behn and Y. C. Ho, "On a Class of Linear Stochastic Differential Games," IEEE Trans. on Automatic Control, Vol. AC-13, pp. 227-240, June 1968.

3. Y. C. Ho, "Optimal Terminal Maneuver and Evasion Strategy," J. SIAM Control, Vol. 4, pp. 421-428, Aug. 1966.

4. Y. C. Ho, "Survey Paper: Differential Games, Dynamic Optimization, and Generalized Control Theory," Journal of Optimization Theory and Applications, Vol. 6, pp. 179-209, Sept. 1970.

5. I. B. Rhodes and D. G. Luenberger, "Differential Games with Imperfect State Information," IEEE Trans. on Automatic Control, Vol. AC-14, pp. 29-38, Feb. 1969.

6. I. B. Rhodes and D. G. Luenberger, "Stochastic Differential Games with Constrained State Estimation," IEEE Trans. on Automatic Control, Vol. AC-14, pp. 476-481, Oct. 1969.

7. W. W. Willman, "Formal Solutions for a Class of Stochastic Pursuit-Evasion Games," IEEE Trans. on Automatic Control, Vol. AC-14, pp. 504-509, Oct. 1969.

8. E. Lehmann, Notes on the Theory of Estimation, University of California Press, Berkeley, California, 1966.

9. M. Mintz, "A Kalman Filter as a Minimax Estimator," Journal of Optimization Theory and Applications, Vol. 9, pp. 99-111, Feb. 1972.

10. D. G. Luenberger, Optimization by Vector Space Methods, Wiley, New York, 1969.

ACKNOWLEDGMENT

The authors wish to express their gratitude for the following support. This work was supported in part by the following agencies and institutions:

By the U. S. Air Force under Grant AFOSR-73-2570, and the Joint Services Electronics Program under Contract DAAB-07-72-C-0259 with the Coordinated Science Laboratory, University of Illinois, Urbana, Illinois.

ON-LINE REGULARIZATION IN INVERSE PROBLEMS

R. RUTMAN
School of Engineering
Tel-Aviv University
Tel-Aviv, Israel

ABSTRACT

The relation between regularization and filtering is considered. Tikhonov's regularization is discussed, and it is shown that it corresponds to physically unrealizable filtering. On-line regularization corresponding to physically realizable filtering is suggested. A procedure for finding regularization parameters based on given filters parameters, and vice versa, is suggested.

I. Problem Statement

We consider identification problems which may be reduced to solving an integral equation of the first type:

$$\int_{-\infty}^{t} r(t-\tau)z(\tau)d\tau = y(t) \qquad (1)$$

for the function z(t). The following problems may be described by such an equation:

a. Signal recovery, z(t) being the signal to be recovered.

b. Identification of the transient response z(t) of a linear time-invariant process, in which case y(t) is the output and r(t) is the input of the process.

c. Determination of sensitivity [5], in which case z(t) is a semi-logarithmic sensitivity function and r(t) is a convolution of the signal at a sensitivity point [6] with the output of the system, etc.

The following considerations must be taken into account:

(1) solving equation (1) is an ill-posed problem as defined in [1,2].

(2) the given data is measured with errors.

(3) the values of the function z(t), in some cases, are to be referred to the same time scale and origin as the given data, r(t) and y(t).

In case "b" this is for predicting the output of the process in control systems using a model of the process; in case "c" this is for calculating the components of gradient in self-optimizing systems. The procedure of solving equation (1) must therefore take place on-line, i.e., the values of the solution are to be obtained on-line. It means that the values of r(t) and y(t) only for $-\infty < t \leq t_1$ must be used when calculating $z(t_1)$, unlike in Tikhonov's regularization where the values of the signals are assumed to be available for all $-\infty < t < \infty$.

We can thus state the following problem. Given equation $Rz = y$, where the linear operator $R \in P$,

$$P\theta = \int_{-\infty}^{t} p(t-\tau)\theta(\tau)d\tau ,$$

and assuming that for some function $\bar{y}$ a solution $\bar{z}$ exists, $\bar{z} = B\bar{y}$, $B \in P$. Find the operator $W \in P$, which is defined for all $y \in L_2(-\infty,\infty)$, such that:

(1) for every ϵ, there exists a δ such that if $\|y(t) - \bar{y}(t)\| < \delta$ it follows that $\|z(t) - \bar{z}(t)\|_{L_2} \leq \epsilon$;

(2) $w(t-\tau) = 0$ for $\tau > t$.

The operator W can be called an operator of on-line regularization.

Taking the above into account, we shall consider in Section II a regularization procedure as defined by A. N. Tikhonov. On-line regularization procedures which differ from Tikhonov's will be considered in the remaining sections.

Let $A(s)$ be the Laplace transform of a time-function $a(t)$. On-line procedure implies physical realizability of a filter with the transfer function $W(s)$. If $W(s)$ is a rational function, it proves the possibility of realization of an on-line procedure by means of a filter with lumped components.

II. Tikhonov's Regularization

Consider the Tikhonov's regularization procedure, which consists in finding a function $z(t)$ minimizing the functional

$$M^{\alpha}[z(\tau), y(t)] = \int_{-\infty}^{\infty} [\int_{-\infty}^{t} r(t-\tau)z(\tau)d\tau - y(t)]^2 dt + \alpha \int_{-\infty}^{\infty} \sum_{k=0}^{p} \phi_k \left(\frac{d^k z(t)}{dt^k}\right)^2 dt \qquad (2)$$

with α and ϕ_k positive constants. An additional functional

$$\alpha \int_{-\infty}^{\infty} \sum_{k=0}^{p} \phi_k \left(\frac{d^k z(t)}{dt^k}\right)^2 dt = \Omega(z)$$

is a Tikhonov's regularizer.

If $\alpha \to 0$, the function $z(t) = z^{\alpha}(t)$ converges uniformly to the exact solution $z(t)$.

V. V. Arsenin and V. V. Ivanov [3,4] have found that

$$W(s) = \frac{R(-s)}{R(s)R(-s) + \alpha\Phi(s)} \qquad (3)$$

where

$$\Phi(s) = \sum_{k=0}^{p} (-1)^k s^{2k} \Phi_k .$$

Since the denominator of W(s) is an even function of s, the corresponding filter is physically unrealizable.

III. On-line Regularization

On proceeding from the Tikhonov's regularization functional, we consider a problem of finding the function z(t), which provides a minimum to the functional M^{α} with an additional condition of being on-line with y(t). Such a solution differs from Tikhonov's one. We shall call the corresponding procedure, an on-line regularization procedure.

Having written (2) according to Parseval's formula as

$$M^{\alpha} = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} Y(s)Y(-s)([R(s)W(s) - 1][R(-s)W(-s) - 1] + + \alpha\Phi(s)W(s)W(-s))ds , \qquad (4)$$

we shall try to find a minimum with an additional condition of physical realizability of W. The necessary and sufficient condition of the minimum of M^{α} is that the first variation equals zero:

$$\delta M^{\alpha} = \frac{1}{\pi j} \int_{-j\infty}^{j\infty} Y(s)Y(-s)([R(s)W(s)-1]R(-s) + \alpha\Phi(s)W(s))\delta W(-s)ds=0. \qquad (5)$$

That means that the integrand must be analytic in the left-half s plane. Thus the transfer function of the filter is

$$W(s) = \frac{1}{[Y(s)Y(-s)]^{+}[R(s)R(-s)+\alpha\Phi(s)]^{+}} \left\{\frac{[Y(s)Y(-s)]^{+} R(-s)}{[R(s)R(-s)+\alpha\Phi(s)]^{-}}\right\}_{+} . \qquad (6)$$

Upper and lower indices "+" or "-" of the rational function A(s) mean, respectively, operations of factorization and splitting [7]:

$$A(s) = [A(s)]^{+}[A(s)]^{-} \; ; \; A(s) = \{A(s)\}_{+} + \{A(s)\}_{-} ,$$

the functions $[A(s)]^{+}$ and $\{A(s)\}_{+}$ containing all the poles of A(s) lying in the left-half s plane, and $[A(s)]^{-}$ and $\{A(s)\}_{-}$ containing all the poles in the right-half s plane; the left and the right zeros of A(s) being divided between $[A(s)]^{+}$ and $[A(s)]^{-}$ respectively.

Let us impose an additional condition on W, namely that it is of the form

$$W(s) = B(s)H(s) , \qquad (7)$$

H(s) being a physically realizable function, $B(s) = R^{-1}(s)$.

If (7) is put in (4) and minimization with respect to H(s) is carried out, one obtains

$$H(s) = \frac{1}{[Y(s)Y(-s)]^{+}[B(s)B(-s)\alpha\Phi(s) + 1]^{+}} \left\{ \frac{[Y(s)Y(-s)]^{+}}{[B(s)B(-s)\alpha\Phi(s) + 1]^{-}} \right\}_{+} , \qquad (8)$$

and the procedure of finding z(t) is divided into two steps: first, filtering the signal y(t) by the filter H(s); second, applying the operation $B = R^{-1}$ to the result.

Such a treatment of the signal y(t) corresponds to an on-line regularization procedure, i.e., the function z(t) minimizes the smoothing functional M^{α} of A. N. Tikhonov with the additional condition formulated above.

We may also try to find a polynomial $\alpha\Phi(s)$, satisfying (8) with H(s) given. Denote

$$[Y(s)Y(-s)]^{+} = \frac{C(s)}{F(s)} , \quad H(s) = \frac{G(s)}{D(s)} , \quad B(s) = \frac{I(s)}{V(s)},$$

$$1 + B(s)B(-s)\alpha\Phi(s) = \Psi(s) \qquad (9)$$

polynomials C(s) and F(s), G(s) and D(s), I(s) and V(s) being mutually simple in pairs.

Thus, instead of (8), one can have a system of polynomial equations [8]

$$UD\Pi^{+} - \Psi^{+}CG = 0 , \qquad (10)$$

$$U\Psi^{-} + NF = C\Pi^{-} , \qquad (11)$$

$$\Gamma\alpha\Phi + \Pi = \Psi . \qquad (12)$$

The solution of this system will give us the unknown polynomial $\alpha\Phi(s)$, as well as polynomials U(s), $\Psi(s)$ and N(s).

The order of the polynomial $\alpha\Phi(s)$, which is twice the order p of the regularizer, may be calculated without solving the system (10) - (12). Denote the order of a polynomial A(s) as l_A, and the index of a rational function A(s) (i.e., the difference between orders of denominator and numerator) as i_A. Using properties of minimal solutions of polynomial equations one may obtain

$$p = i_Y + i_H - i_R - 1 . \qquad (13)$$

The solution of the homogeneous polynomial equation (10), according to [8], may be written in the form

$$U = U_0 CG , \qquad (14)$$

$$\Psi^{+} = \Psi_0^{+} D\Pi^{+} , \qquad (15)$$

$U_0 = \Psi_0^+$ being an arbitrary polynomial. Having introduced this solution into (11), we obtain:

$$U_0(s)U_0(-s)G(s)D(-s) + N_0F(s) = 1 \tag{16}$$

and

$$N(s) = N_0(s)C(s)\Pi^-(s), \quad l_{U_0} = l_F - l_C - l_G - 1 .$$

Having compared the number of unknown quantities with the number of algebraic equations which are created by polynomial equations (14) and (12), we obtain the necessary condition of solvability of the system:

$$\gamma \geq l_G + l_I + l_C \tag{17}$$

where γ is the number of unspecified parameters in the polynomials $F(s)$, $D(s)$ and $G(s)$. We obtain another necessary condition from (13):

$$i_\psi + i_H \geq i_R + 1 . \tag{18}$$

IV. Examples

We will consider approximate differentiation of the signal $y(t) = e^{-at}$, i.e., $r(t-\tau) = 1$, $R(s) = 1/s$, $B(s) = s$, $C(s)/F(s) = 1/(s+a)$. Let us first take

$$W(s) = sg/(1 + ds) \tag{19}$$

($d > 0$), so that

$$H(s) = g/(1 + ds) , \tag{20}$$

and

$$l_D = 1, \; l_G = 0, \; i_H = 1, \; i_I = 1, \; l_V = 0, \; i_R = 1 . \tag{21}$$

Equation (18) holds and (17) becomes $\gamma \geq 1$. We therefore do not specify the coefficient g.

According to (13), $p = 0$. From (15):

$$\Psi^+(s) = u_0(1 + ds) ; \tag{22}$$

hence

$$1 - s^2\alpha\Phi(s) = u_0^2 - u_0^2d^2s^2 \tag{23}$$

and

$$u^2 = 1, \quad \alpha = d^2 .$$

We obtain from (16):

$$g = \frac{1}{1 + ad} . \tag{24}$$

The approximate differentiation by means of filter (19) therefore corresponds to on-line regularization, the regularizer being of zero order:

$$\Omega(z) = \alpha \int_{-\infty}^{\infty} z^2(t)dt , \tag{25}$$

$\alpha = d^2$.

If instead of (20)

$$H(s) = \frac{1}{(s+d_1)(s+d_2)} \tag{26}$$

is given ($d_1 > 0$, $d_2 >$), we can prove that

$$\Omega(z) = \alpha \int_{-\infty}^{\infty} [\phi_0 z^2 + (\frac{dz}{dt})^2]dt , \tag{27}$$

α being $1/d^2 d_2^2$ and $\phi_0 = d_1^2 + d_2^2$. (28)

V. Conclusions

One-to-one correspondence is established between on-line regularization and filtering. For r(t) and y(t) stochastic processes with known power spectra, the resulting filtering problem is the Wiener problem. Similarly, Tikhonov's regularization procedure corresponds to unrealizable Wiener filter. This is the reason for the presence of Y(s) and R(s) in the expression (6) for W(s). Similar a priori information is needed in order to choose the value of α and the type of Tikhonov's regularizer. Conversely, with α and the type of regularizer specified, the procedure of regularization is optimal only with respect to specified input data. The established relations make the parallelism of the filtering and the regularization methods apparent.

REFERENCES

1. A. N. Tikhonov. On the Solution of Ill-Posed Problems and the Regularization Method. Doklady A. K. Nauk, vol. 151, No. 3, 1963.

2. A. N. Tikhonov. On the Regularization of Ill-Posed Problems. Doklady Ak. Nauk, vol. 153, No. 1, 1963.

3. V. Ya. Arsenin, V. V. Ivanov. On the Solution of Some Convolution Type Integral Equations of the First Kind by means of the Regularization Method. Zhurnal Vychislitelnoy Matematiki i Mat. Fiziki, vol. 8, No. 2, 1968.

4. V. Ya. Arsenin, V. V. Ivanov. On the Resolution and Accuracy of the Regularization Method for Solving Convolution Type Integral Equations of the First Kind. Radiotekhnika i Elektronika, vol. 14, No. 1, 1969.

5. R. S. Rutman. The Method of Three Points in Sensitivity Theory. Engineering Cybernetics, No. 4, 1968.

6. P. Kokotović. Method of Sensitivity Points in the Investigation and Optimization of Linear Control Systems. Automation and Remote Control, vol. 25, No. 12.

7. S. S. L. Chang. Synthesis of Optimum Control Systems. McGraw-Hill, 1961.

8. L. N. Volgin. Elements of the Theory of Control Computers, Sovietskoye Radio, 1962.

DECENTRALIZED STOCHASTIC CONTROL OF AN INTERCONNECTED ELECTRIC ENERGY SYSTEM*

EDGAR C. TACKER**
Departments of Engineering Science, Electrical Engineering, and
Chemical Engineering, Louisiana State University at Baton Rouge

CHARLES W. SANDERS
Electrical & Computer Engineering Department, The University of
Wisconsin at Madison

DAVID R. FISCHER
Electrical Engineering Department, Louisiana State University at
Baton Rouge

THOMAS D. LINTON
Data Acquisition & Simulation Laboratory, The University of
Wisconsin at Madison

ABSTRACT

Recent theoretical results of the authors in decentralized filtering and control are applied to the problem of designing a controller for a two-area electric energy system. Performance comparisons show that the resulting decentralized stochastic controller, while much simpler in structure, competes favorably with a centralized optimal controller.

INTRODUCTION

Interconnected electric energy systems provide an important example of a class of large-scale systems which are informationally decentralized. This is not always a crucial factor when the interconnection network is small, and in these cases centralized control techniques can be used quite effective (e.g., see references (1,2)). However, in the more general case the communications/computational costs involved in implementing a centralized controller often become prohibitive, and decentralization of some sort becomes manditory. Any scheme for controlling such systems must pay particular attention to the information flow constraints present in the system.

PROBLEM STATEMENT

In this paper a control scheme based on recent (3,4) theoretical work in decentralized filtering and control is developed. This method is then used to develop a decentralized controller for a 2-area electric power system. The model used for the power system is of the form (see (1,2) for further modeling details):

---

* This research is sponsored by The Air Force Office of Scientific Research , US Air Force Systems Command under Contract F44620-68-C-0021. The third named author received support from the National Science Foundation in the form of an NSF Traineeship.

** Presently--Visiting Associate Professor, Electrical & Computer Engineering Department, The University of Wisconsin at Madison.

Area S_i

State Variables:	Δf_i	Frequency deviation
	Δp_{ti}	Turbine output deviation
	Δs_{gi}	Governor position deviation
Output Variable:	Δf_i	
Control Variable:	Δp_{ci}	Speed changer increment
Interaction Variable:	$u_i = \Delta p_{tie,i}$	Tie-line power flow

(where $u_1 = -u_2$ for a 2-area power system).

The composite state equation is of the form

$$\dot{x} = Ax + B_1 m_1 + B_2 m_2 + Ld$$

where $x = (x_1, x_2)'$, $d = (d_1, d_2)'$, and d_i represents "load disturbance" in S_i.

THE DECENTRALIZED CONTROLLER

The decentralized controller consists of two local filter/controllers FC1 and FC2, wherein FCi is the cascade of the local filter $\mathcal{F}_i$ and local controller $\mathcal{C}_i$.

Local Filter: The local filter is complexity-constrained to be of the form

$$\mathcal{F}_i: \quad \dot{\hat{x}}_i = E_i \hat{x}_i + G_i y_i$$

wherein $\hat{x}_i$ is an unbiased estimate of x_i. As shown in (3) unbiasedness, together with another very reasonable restriction on the class of admissible filters leads to a deterministic parameter optimization problem. The resulting filtering algorithm (3) is very similar to the standard Kalman algorithm--projected onto the subspace of the local state, and reduces to the Kalman form as the interaction coupling is decreased to zero.

Local Controller: The local controller is also complexity-constrained, and the following output-feedback form is used

$$\mathcal{C}_i: \quad m_i = -F_i \hat{z}_i$$

where $\hat{z}_i \triangleq (\hat{x}_i, \widehat{IACE}_i)'$, $\widehat{ACE}_i \triangleq b_i \widehat{\Delta f}_i + u_i$, $\widehat{IACE}_i \triangleq \int \widehat{ACE}_i \, dt$,

and b_i is the industry specified "tie-line bias constant" (1). Reference (4) gives two computational techniques for numerically generating the optimal feedback matrices F_1^* and F_2^*.

PERFORMANCE COMPARISONS

The resulting decentralized controller is seen to be considerably simpler in structure than the overall optimal centralized stochastic controller. Preliminary numerical studies indicate that the performance (using the cost functional defined in (1,2)) of the decentralized controller compares quite favorably with the centralized optimal controller. Further research is needed in order to more fully evaluate the true potential of such decentralized structures relative to future control system implementations.

REFERENCES

1. Tacker, E.C., Reddoch, T.W., Tan, O.T., and Linton, T.D., "Automatic Generation Control of Electric Energy Systems--A Simulation Study", IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-3, No. 4, July, 1973.

2, Tacker, E.C., Lee, C.C., Reddoch, T.W., Tan, O.T., and Julich, P.M., "Optimal Control of Interconnected Electric Energy Systems--A New Approach", Proc. IEEE, Vol. 60, No. 10, Oct., 1972.

3. Sanders, C.W., Tacker, E.C., and Linton, T.D., "A Decentralized Filter For Interacting Dynamical Systems", Proc. 4th Symposium on Nonlinear Estimation and Its Applications, San Diego, Calif., Sept., 1973.

4. Linton, T.D., Fischer, D.R., Tacker, E.C., and Sanders, C.W., "Decentralized Control of an Interconnected Electric Energy System Subject to Information Flow Constraints", 1973 IEEE Conference on Decision & Control, San Diego, Calif., Dec., 1973 (To appear).

THE INTERRELATED TREE ANATOMIES OF LADDERS, WHEELS, DRUMS AND TOPS

BASIL R. MYERS
Department of Electrical Engineering
University of Notre Dame, Notre Dame, Indiana 46556

Abstract - An analysis is presented which reveals the interrelationships among the numbers of spanning forests of paths and cycles and of the spanning trees of ordinary and multigraph ladders, wheels, drums and tops, and which simultaneously reveals the interrelationships between these numbers, the Fibonacci and Lucas numbers, and the binomial coefficients.

INTRODUCTION

The join $G_1 + G_2$ of two graphs G_1 and G_2 is the graph obtained on adjoining each vertex in G_1 to each vertex in G_2, by a single edge. For example, the complete bipartite graph $K_{m,n} = \overline{K}_m + \overline{K}_n$, where $\overline{K}_x$ denotes a set of x isolated vertices and is the complement [1] of the complete graph K_x of order x, which is an ordinary graph with every pair of its vertices adjacent. A *simple* ("terminated") *ladder* L_{n+1} of order n + 1 is the join $K_1 + P_n$ of K_1 and the path P_n of order n; and the *wheel* W_{n+1} of order n + 1 is the join $C_n + K_1$ of K_1 and the cycle C_n of order n, where C_n is called the rim of W_{n+1}, K_1 is the hub, and the n edges incident with K_1 are called spokes. Clearly, if r is a rim edge of W_{n+1}, then $W_{n+1} - r = L_{n+1}$.

The cartesian product $G_1 X G_2$ of G_1 and G_2 is defined by considering any two points $u = (u_1,u_2)$ and $v = (v_1,v_2)$ in the product $V_1 X V_2$ of the respective vertex sets V_1, V_2 of G_1,G_2. Then u and v are adjacent in $G_1 X G_2$ whenever $[u_1 = v_1$ and u_2 adj $v_2]$ or $[u_2 = v_2$ and u_1 adj $v_1]$. The ("terminated") *parallel ladder* M_{2n} of order 2n is the product $P_n \times P_2$ of the paths P_n, P_2; and the *drum* D_{2n} of order 2_n is the product $C_n X P_2$ of the cycle C_n and the path P_2. Clearly, M_{2n} is a subgraph of D_{2n}.

The *modified parallel ladder* M'_{2n} of order 2n is obtained by adjoining an isolated vertex K_1 to each of the n odd-numbered vertices in sequence in the path P_{2n-1} of order 2n-1. There is one-to-one correspondence between the spanning forests of M'_{2n} and those of M_{2n} [5], and it happens to be more convenient to enumerate the forests of M'_{2n} rather than those of M_{2n}.

Graphs L_5, W_5, M_8, M'_8 and D_8 are shown in Figs. 1 (a)-(e), respectively. It is easy to see that each of L_{n+1}, W_{n+1}, M_{2n}, M'_{2n} and D_{2n} is a planar graph so that each has a planar dual. The duals of L_{n+1}, M_{2n} and M'_{2n} are multigraph ladders, and the ordinary wheel W_{n+1} is its own self-dual. For the drum D_{2n}, the dual graph is the *top* S_{n+2}, illustrated by S_5 in Fig. 1(f). It is well known that a planar graph has the same number of spanning trees as its dual and that it is often more convenient in tree enumeration to work with the dual rather than the original. An example of this will be given later in computing the trees of D_{2n} from those of S_{n+2}.

The literature is replete in techniques for enumerating the spanning forests and trees of ladders and wheels [3]-[10], though the drum D_{2n} and its dual, the top S_{n+2}, do not appear to have attracted particular attention. It is known, for instance, that the number $T(L_{n+1})$ of the spanning trees of the ordinary, simple (terminated) latter L_{n+1} is [4]

$$T\ (L_{n+1}) = F_{2n} \tag{1}$$

where F_k is the k-th number in the Fibonacci series $F_0, F_1, F_2, F_3, \ldots, F_k, \ldots = 0, 1, 1, 2, \ldots, F_k = F_{k-1} + F_{k-2}, \ldots$.

It is also known that the number $T\ (W_{n+1})$ of the spanning trees of the ordinary wheel W_{n+1} is [7]

$$T\ (W_{n+1}) = L_{2n} - 2 \tag{2}$$

where L_k is the k-th number in the Lucas series $L_1, L_2, L_3, \ldots, L_k, \ldots = 1, 3, 4, \ldots, L_k = L_{k-1} + L_{k-2} = F_{k-1} + F_{k+1}$. Bedrosian and others [6] [8] [9] have extended these results to certain multigraph wheels.

In any event, it is evident that these several classes of graphs have certain interrelated topological anatomies, these interrelationships involving Fibonacci (or Lucas) number structures. The purpose of the present paper is to report an analysis of these numerical interrelationships. It is found that tables can be constructed which reveal the common features of their several tree counts in terms of the numbers of spanning forests of the cycle or path subgraphs which appear as factors in the join or product definitions given earlier.

PASCAL'S TRIANGLE AND THE FIBONACCI AND LUCAS SEQUENCES

Of central interest is the known [2] table, given in part in Table I, which shows the relationship between the Fibonacci numbers and the binomial coefficients (the coefficients $\binom{n}{k}$ of x^k in the expansion of $(1 + x)^n$). The column sums are the Fibonacci numbers, as indicated, and Pascal's triangle of the binomial coefficients (1, 121, 1331, etc.) can be recognized, as shown. The integer $I_{i,j}$ in row i and column j of the table is given by the triangular relationship $I_{i,j} = I_{i,j-2} + I_{i-1,j-1}$ (e.g., 14 = 5 + 9 is the integer in row 4, column 7).

TABLE 1: The Fibonacci Sequence and Pascal's Triangle

Column sums:	F_0	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
		1		1		1		1	
	0		1		2		3		4
		0		1		3		6	
			0		1		4		10
				0		1		5	
					0		1		6
						0		1	
							0		1
								0	
									0

The sums of the integers in the column F_{2n} thus gives the number $T(L_{n+1})$ of spanning trees in the simple ladder L_{n+1}. For example,

$$T\ (L_6) = 5 + 20 + 21 + 8 + 1 = 55 \tag{3}$$

which is a 5-part partition of 55 into the parts 5, 20, 21, 8 and 1. It will be shown that these parts are actually weighted sums of the spanning forests in the path P_n defined by $L_{n+1} = K_1 + P_n$, and, subsequently, that the numbers of the spanning trees of W_{n+1}, M_{2n}, D_{2n} and S_{n+2} similarly partition into parts which are weighted sums of the spanning forests either of C_n or of P_n.

The structural concept behind Table 1 leads naturally to a similar table, Table 2, for the Lucas numbers.

TABLE 2: The Lucas Sequence

Column sums:	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8
		2		2		2		2
	1		3		5		7	
		1		4		9		16
			1		5		14	
				1		6		20
					1		7	
						1		8
							1	
								1

The table shows how the Lucas numbers relate to the coefficients of x^k in the expansion of $(1 + 2x)\ (1 + x)^{n-1}$. Since the number $T(W_{n+1})$ of spanning trees of the wheel W_{n+1} is $L_{2n}-2$, the set of integers which sum to L_{2n} in the respective column in Table 2 thus identifies a partition of the integer $T(W_{n+1})$; for example,

$$T\ (W_6) = 25 + 50 + 35 + 10 + 1 = 121 \tag{4}$$

It will be shown that these parts of $T(W_{n+1})$ are weighted sums of the spanning forests of the cycle C_n in the identity $W_{n+1} = K_1 + C_n$.

TREE ANATOMIES OF ORDINARY LADDERS AND WHEELS

Denote by $t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \dots t_{\alpha_j}^{k_j}$ the number of spanning forests (k-trees) of a connected graph G_n of order n, each of which has k_i (>0) components containing a_i(>0) of the n vertices of G_n, $i = 1, 2, \dots, j$, so that $k = k_1 + k_2 + \dots + k_j$ is the number of components in each forest, and $n = k_1\alpha_1 + k_2\alpha_2 + \dots + k_j\alpha_j$. There are thus as many different numbers $t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \dots t_{\alpha_j}^{k_j}$ as there are partitions of n. For example, the integer 3 has the three partitions 3, 2 + 1, and 1 + 1 + 1, the corresponding k-tree terms being t_3, t_2t_1, and t_1^3, respectively.

The simply ordinary ladder L_{n+1} is the join $K_1 + P_n$ of the isolated vertex K_1 and the path P_n, so that K_1 has degree n in L_{n+1}. Suppose $t(L_{n+1})$ is a spanning tree of L_{n+1}, so that $t(L_{n+1})$ necessarily contains d, where $1 \leq d \leq n$, of the n edges incident with the vertex K_1 in L_{n+1}. Further $t(L_{n+1}) - K_1$ is a tree of P_n when $d = 1$ and is a spanning forest (k-tree), with $k > 1$ components, of P_n otherwise. That is, $t(L_{n+1}) - K_1$ is in any event one of the spanning forests of P_n whose number is $t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \dots t_{\alpha_j}^{k_j}$. It is easy to see that each component t_i' (itself a tree) in each such k-tree of P_n must be connected to K_1, in $t(L_{n+1})$, by exactly one edge.

Suppose that t_i' is of order α_i (and hence that these α_i vertices are adjacent to K_1 in L_{n+1}) and that there are k_i such components t_i' each of this order. There are α_i ways of joining each of these components to K_1 by a single edge, towards forming the tree $t(L_{n+1})$ of G, and hence $\alpha_i^{k_i} t_{\alpha_i}^{k_i}$ such ways of joining the $t_{\alpha_i}^{k_i}$ components to K_1. It follows immediately that there are

$$(\alpha_1^{k_1} \alpha_2^{k_2} \dots \alpha_j^{k_j}) (t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \dots t_{\alpha_j}^{k_j})$$

$$= t_{(\alpha),k} (L_{n+1}), \text{ say,} \tag{5}$$

ways of connecting the $t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \dots t_{\alpha_j}^{k_j}$ forests of P_n to K_1 to form trees of L_{n+1}. Thus, the total number $T(L_{n+1})$ of the spanning trees of L_{n+L} partitions into the n integers $t_{(\alpha),k} (L_{n+1})$ corresponding to the respective k-tree numbers in P_n. For example, for n = 4 the quantities

$t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \cdots t_{\alpha_j}^{k_j}$ are the binomial coefficients $t_4 = 1$, $(t_3t_1 + t_2^2) = 2 + 1 = 3$, $t_2t_1^2 = 3$, and $t_1^4 = 1$; and the respective partitions of $T(L_5)$ are $t_{(\alpha),1} = 4.1 = 4$, $t_{(\alpha),2} = 3.2 + 4.1 = 10$, $t_{(\alpha),3} = 2.3 = 6$, $t_{(\alpha),4} = 1$, so that $T(L_5) = 21$ (which is the Fibonacci number F_8). It is observed that these numbers $t_{(\alpha),k}\ (L_{n+1})$ are precisely those in the F_8 column of Table 1 (excluding the zero entry in each column). It can be shown that this is true in general, as stated in the following theorem (given here without proof, because of space limitations):

Theorem. The number $T(L_{n+1})$ of the spanning trees of $L_{n+1} = K_1 + P_n$ is the weighted sum

$$T\ (L_{n+1}) = \sum_{k=1}^{n} t_{(\alpha),k}\ (L_{n+1}) \tag{6}$$

of the numbers of k-trees (spanning forests) of P_n.

In similar fashion, it can be shown that:

Theorem. The number $T(W_{n+1})$ of the spanning trees of $W_{n+1} = K_1 + C_n$, plus 2, is the weighted sum

$$2 + T\ (W_{n+1}) = \sum_{k=1}^{n} t_{(\alpha),k}\ (W_{n+1}) \tag{7}$$

of the numbers of k-trees (spanning forests) of C_n.

Analysis of the modified parallel ladder M'_{2n}, whose spanning forests are in one-to-one correspondence with those of the ordinary ladder M_{2n}, is equally straightforward. M'_{2n} is obtained by adjoining an isolated vertex K_1 to each of the n odd-numbered vertices in sequence in the path P_{2n-1} of order 2n-1. For this structure, the spanning forests whose number is $t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \cdots t_{\alpha_j}^{k_j}$ are those defined on the n odd-numbered vertices of P_{2n-1}. Whereas deleting k edges of a path P_n of order n creates a (k+1)-tree of P_n, it is easy to see that there are 2^k ways of deleting edges in P_{2n-1} to create a (k+1)-tree of P_{2n-1} with respect to its n odd-numbered vertices. Thus, the number of weighted spanning forests $t_{(\alpha),k}\ (M'_{2n})$ is given by

$$t_{(\alpha),k}\ (M'_{2n}) = 2^{k-1}\ t_{(\alpha),k}\ (L_{n+1}) \tag{8}$$

For example, since (from Table 1)

$$T\ (L_5) = 4 + 10 + 6 + 1 = 21 \tag{9}$$

then

$$T\ (M'_8) = 2^0.4 + 2^1.10 + 2^2.6 + 2^3.1 = 56 = T(M_8) \tag{10}$$

is the number of spanning trees of M'_8 and of M_8.

A closed formula for $T(M_{2n})$, from determinantal methods, has been reported earlier by Sedláček (see Moon [11]). The present result, equ. (8), now shows how $T(M_{2n})$ partitions in terms of weighted sums of spanning forests of the path subgraph P_{2n-1} in M'_{2n}, factors of the form 2^{k-1} relating these sums to the Fibonacci numbers.

TREE ANATOMIES OF THE DRUM AND TOP

The drum $D_{2n} = C_n \times P_2$ of order $2n$ is a planar graph which has the top S_{n+2} as its geometric dual. Thus, the number $T(D_{2n})$ of trees of D_{2n} is equal to the number $T(S_{n+2})$ of those of S_{n+2}, and there is a one-to-one correspondence between their trees. Suffice it, therefore, to consider only S_{n+2} which is the graph obtained on adjoining each vertex in the cycle C_n of the wheel $W_{n+1} = C_n + K_1$ to an isolated vertex K'_1, so that $S_{n+2} - K_1 - K'_1 = C_n$. It will be shown that the number $T(S_{n+2})$ partitions into a sum of n parts corresponding to the respective k-trees, $k = 1, 2, \ldots, n$, of the subgraph C_n, and that these parts can be obtained directly from the Lucas sequence table, Table 2.

As previously, denote the number of k-tress of C_n by $t_{\alpha_1}^{k_1} t_{\alpha_2}^{k_2} \ldots t_{\alpha_j}^{k_j}$ where $k = k_1 + k_2 + \ldots + k_j$ and $n = k_1\alpha_1 + k_2\alpha_2 + \ldots + k_j\alpha_j$. Suppose $t(S_{n+2})$ is a spanning tree of S_{n+2} which has one of these k-trees of C_n as a subgraph. There is necessarily a unique path P joining the vertices K_1 and K'_1 in $t(S_{n+2})$, this path containing m, say, of the n vertices of C_n, where $m > 0$. The sub-path $P' = P - K_1 - K'_1$ of length $m-1 \geq 0$ is a path in some component t'_i of the k-tree of C_n in question. Suppose t'_i has α_i of the n vertices of C_n. If there are k_i such components t'_i, then there are clearly $k_i\alpha_i^2$ ways of forming P so that P' is in one of these k_i components; and $2\alpha_j$ ways of connecting a component t'_j of the k-tree which, in $t(S_{n+2})$, does not contain P'. It is then not too difficult to show that there are

$$n. \, 2^{k-1} \cdot \prod_{i=1}^{j} \alpha_i^{k_i} \tag{11}$$

ways of forming a tree $t(S_{n+2})$ and hence that:

Theorem. The number of spanning trees of S_{n+2} (and of D_{2n}) is

$$T(S_{n+2}) = n. \sum_{(\alpha)} 2^{k-1} \cdot \prod_{i=1}^{j} (\alpha_i t_{\alpha_i})^{k_i} \tag{12}$$

where the sum (α) is over all partitions of the form $n = k_1\alpha_1 + k_2\alpha_2 \ldots + k_j\alpha_j$.

The factor n in equ. (12) is significant since it shows that the complexity of computation of $T(S_{n+2})$ is far less marked, by virtue of this factor, than would otherwise be the case. In any event, $T(S_{n+2})/n$ is seen to partition into the n parts

$$t_{(\alpha),k}(S_{n+2}) = 2^{k-1} \cdot \prod_{i=1}^{j} (\alpha_i t_{\alpha_i})^{k_i} \qquad (13)$$

where $k = 1, 2, \ldots, n$, and where each of these parts is determined by the corresponding weighted sum of the k-trees of C_n.

For the top S_5 of Fig. 1(f), for example, the quantities $t_{(\alpha),k}$ involved in computing $T(S_5)$ from equ. (12) are as follows:

C_3 k-trees	k	2^{k-1}	$\Pi(\alpha_i t_{\alpha_i})^{k_i}$	$t_{(\alpha),k}$
$t_3 = 3$	1	1	9	9
$t_2t_1 = 3$	2	2	6	12
$t_1^3 = 3$	3	4	1	4
				25

There are thus 3.2 = 75 spanning trees in S_5, and in its dual graph, the drum D_6. The numbers $\Pi(\alpha_i t_{\alpha_i})^{k_i}$ are precisely the numbers in the second through last positions in column L_{2n} of the Lucas sequence table, Table 2.

MULTIGRAPHS

Extension of the previous analysis to certain multigraph structures follows almost routinely. Consider, for example, the multigraph wheel $W_{n+1,\rho,\nu}$ of order n + 1, with spoke edge multiplicity ρ and rim edge multiplicity ν, so that $W_{n+1,\rho,\nu} = K_{1,\nu} + C_{n,\rho}$, say. The k-trees of $C_{n,\rho}$ are independent of ν, and the respective numbers of the k-trees is simply ρ^{n-k} times the numbers for the ordinary graph, when $\rho = \nu = 1$. The consequence of the spoke multiplicity is clearly to multiply the weighting factor $\alpha_i^{k_i}$ for the ordinary graph by the factor ν^{k_i}.

The following example illustrates these modifications. Ordinary graph W_5:

$$T(W_5) = 4t_4 + 3.1\, t_3t_1 + 2^2 \cdot t_2^2 + 2.1^2 \cdot t_2t_1^2 + 1^4 \cdot t_1^4 \qquad (14)$$

where the k-tree numbers $t_{\alpha_1}^{k_1}\, t_{\alpha_2}^{k_2} \ldots t_{\alpha_j}^{k_j}$ are those for the cycle C_4, viz.;

$t_4 = 4$, $t_3t_1 = 4$, $t_2^2 = 2$, $t_2t_1^2 = 4$, $t_1^4 = 1$.

Multigraph $W_{5,\rho,\nu}$;

$$T(W_5) = (4.\nu)\, u_4 + (3.\nu)(1.\nu)\, u_3u_1 + (2.\nu)^2\, u_2^2 + (2.\nu)(1.\nu)^2\, u_2u_1^2 + (1.\nu)^4\, u_1^4$$

where the k-tree numbers $u_{\alpha_1}^{k_1}\, u_{\alpha_2}^{k_2} \ldots u_{\alpha_j}^{k_j}$ are those for $C_{4,\rho}$, so that $u_4 = \rho^3 t_4 = 4\rho^3$, $u_3u_1 = \rho^2 t_3t_1 = 4\rho^2$, $u_2^2 = \rho^2 t_2^2 = 2\rho^2$, $u_2u_1^2 = \rho t_2t_1^2$, $u_1^4 = t_1^4 = 1$.

Thus,

$$T(W_{5,\rho,\nu}) = 16\nu\rho^3 + 20\nu^2\rho^2 + 8\nu^3\rho + \nu^4 \qquad (15)$$

which, of course, reduces to $T(W_5) = 45$ when $\rho = \nu = 1$. It is seen that ν is a factor of $T(W_{n+1,\rho,\nu})$ and that the general formula for $T(W_{n+1,\rho,\nu})$ is

$$T(W_{n+1,\rho,\nu} = \sum_{k=1}^{n} t_{(\alpha),k}(W_{n+1}).\nu^k\rho^{n-k} \qquad (16)$$

where the $t_{(\alpha),k}(W_{n+1})$ are the parts of $T(W_{n+1})$ which are weighted sums of k-trees of C_n and which can be read from the table of Lucas numbers, Table 2. Previous results by Bose et al. [8] and by Johnson and Johnson [9] have given closed forms for the number $T(W_{n+1,\rho,\nu})$. That due to Johnson and Johnson is particularly interesting:

$$T(W_{n+1,\rho,\nu}) = 2\rho^n[Q_n(1 + \nu/2\rho) -1] \qquad (17)$$

where Q_n is the Chebyshev polynomial of the first kind of degree n.

The analysis for the multigraph ladders, drum and top follows immediately. For the simple ladder $L_{n+1,\rho,\nu} = k_{1,\nu} + P_{n,\rho}$ for example,

$$T(L_{n+1,\rho,\nu}) = \sum_{k=1}^{n} t_{(\alpha),k}(L_{n+1}).\ \nu^k\rho^{n-k} \qquad (18)$$

where the $t_{(\alpha),k}(L_{n+1})$ parts are obtained directly from the Fibonacci sequence table, Table 1. Thus

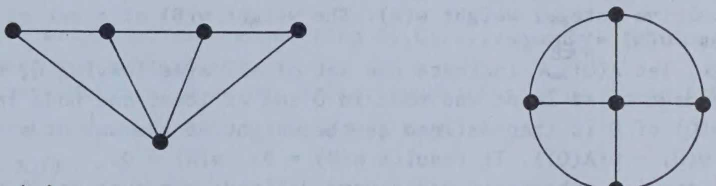
$$T(L_{6,\rho,\nu}) = 5\nu\rho^4 + 20\nu^2\rho^3 + 21\nu^3\rho^2 + 8\nu^4\rho + \nu^5 \qquad (19)$$

which reduces to $T(L_6) = 55$ when $\rho = \nu = 1$.

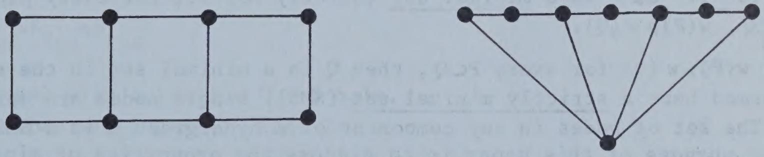
REFERENCES

1. HARARY, F.: "Graph theory" (Addison-Wesley, Reading, Mass., 1969).
2. ALFRED, U.: "An introduction to Fibonacci discovery" (The Fibonacci Association, San Jose, Calif., 1965).
3. MOON, J.W.: "Counting labelled trees" (Canadian Mathematical Monographs, No. 1, Canadian Mathematical Congress, Montreal, 1970).
4. MORGAN-VOYCE, A.M.: "Ladder network analysis using Fibonacci numbers, Trans. IRE, 1959, CT-6, pp. 321-322.
5. MYERS, B.R.: "Number of trees in a cascade of 2-port networks," Trans. IEEE, 1967, CT-14, pp. 284-290.

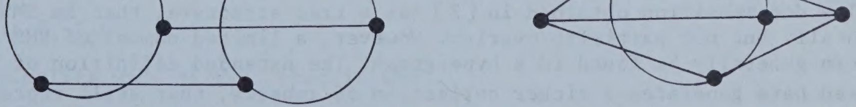
6. BEDROSIAN, S.D.: "Number of spanning trees in multigraph wheels," Trans. IEEE, 1972, CT-19, pp. 77-78.
7. MYERS, B.R.: "Number of spanning trees in a wheel," Trans. IEEE, 1971, CT-18, pp. 280-282.
8. BOSE, N. K., FEICK, R., and SUN, F.K.: "General solution to the spanning tree enumeration problem in multigraph wheels," Trans. IEEE, 1973, CT-20, pp. 69-70.
9. JOHNSON, D.E., and JOHNSON, J.R. (with a reply by BOSE, N.K.): "Comment on 'General solution to the spanning tree enumeration problem in multigraph wheels'", Trans. IEEE, 1973, CT-20, pp. 454-455.
10. HARARY, F., SCHWENK, A.J., O'NEIL, P.V., and READ, R.C.: "The number of trees in a wheel," presented at the 700th Meeting of the Amer. Math. Soc., Case Western Reserve University, Nov. 25, 1972.



(a) Simple ladder L_5 (b) Wheel W_5



(c) Parallel ladder M_8 (d) Modified parallel ladder M_8'



(e) Drum D_6 (f) Top S_5

Fig. 1

HYPERGRAPH DECOMPOSITION WITH MINIMAL CONNECTIONS

E. LODI and F. LUCCIO
University of Pisa, Pisa, Italy

Minimal sets (MS's) of nodes of a weighted hypergraph are defined, with the intuitive meaning of classes of nodes strictly related to one another. Based on the definition, the properties of MS's are discussed; in particular, conditions are derived under which MS's may have nonvoid intersections. Some results are also presented, leading to the generation of new MS's from the union of two or more given MS's.

PRESENTATION OF PROBLEM

As known [1], a loop free hypergraph $H=(N,A)$ is defined as a set of nodes N and a set of arcs A, such that every $a \in A$ is incident to the nodes of a subset $Q \subseteq N$, $\#(Q) \geqslant 2$. If every Q contains exactly two nodes, H reduces to a graph.

In a weighted hypergraph a function $w : A \rightarrow I^{+}$ is defined, assigning to any $a \in A$ a positive integer weight $w(a)$. The weight $w(B)$ of a set of arcs $B \subseteq A$ is defined as $w(B) = \sum_{a \in B} w(a)$.

For $Q \subseteq N$, let $A(Q) \subseteq A$ indicate the set of all arcs leaving Q, that is all arcs incident to at least one node in Q and at least one node in $N-Q$. The weight $w(Q)$ of Q is then defined as the weight of the set of arcs leaving Q, that is $w(Q) = w(A(Q))$. It results $A(N) = \emptyset$; $w(N) = 0$.

In [2] special subsets of nodes were defined; any such set Q had the intuitive meaning of a collection of nodes with internal connections stronger than the connections to nodes in $N-Q$. The basic definition of [2] is extended here as follows:

Definition 1. $Q \subseteq N$ is a minimal set (briefly MS) if, for every proper subset $P \subset Q$, $w(P) \geqslant w(Q)$.

If $w(P) > w(Q)$ for every $P \subset Q$, then Q is a minimal set in the sense of [2], termed here a strictly minimal set (SMS). Single nodes are defined as SMS's. The set of nodes in any component of a hypergraph H is a SMS.

The purpose of this paper is to discuss the properties of minimal sets, and the structure of the relative hypergraph decomposition. This problem is relevant in several areas, from electrical design to information retrieval (for a list of referenced applications, see [3]). In all cases, the mathematical description of the structure under examination is the one of a hypergraph, since different elements of the structure (nodes) may have a common relation (arc incident to such nodes).

The decomposition obtained in [2] has a tree structure, that is SMS's can totally but not partially overlap. However, a limited number of SMS's can generally be found in a hypergraph. The extended definition of MS proposed here generates a richer collection of subsets, that still represent portions of the hypergraph whose nodes are strictly related to one another.

This research was supported by Consiglio Nazionale delle Ricerche, under research grant No 72.00254.42 and a post-graduate fellowship issued by the Comitato Nazionale per le Scienze Matematiche.

It will be shown that the resulting decomposition has more intrigued properties, since MS's can partially overlap under special conditions.

STRUCTURE OF DECOMPOSITION

Some results on MS's are now presented, and the structure of decomposition is investigated. Additional notation is in order [2]. Let $Q \subseteq N$ be a nonvoid subset of nodes; $\pi = \{G_1, G_2, \dots, G_n\}$ be a partition on Q; $G_1, G_2, \dots, G_m$ be a nonvoid collection of sets from among $G_1, G_2, \dots, G_n$. Then:

$$\alpha^{\pi}_{G_1 G_2 \dots G_m} = w[A(G_1) \cap A(G_2) \cap \dots \cap A(G_m) \cap A(G_1 \cup G_2 \cup \dots \cup G_m) - \\ - A(G_1) \cap A(G_2) \cap \dots \cap A(G_m) \cap A(G_{m+1} \cup G_{m+2} \cup \dots \cup G_n)]$$

is the total weight of arcs connecting $G_1, G_2, \dots, G_m$ and $N-Q$ only.

$$\beta^{\pi}_{G_1 G_2 \dots G_m} = w[A(G_1) \cap A(G_2) \cap \dots \cap A(G_m) - \\ - A(G_1) \cap A(G_2) \cap \dots \cap A(G_m) \cap A(G_1 \cup G_2 \cup \dots \cup G_m)]$$

is the total weight of arcs connecting $G_1, G_2, \dots, G_m$ only. (Obviously, β's have at least two subscripts).

Whenever ambiguity does not arise, superscript π in α and β is omitted.

Theorem 1. Let K_1 be a SMS; K_2 be a MS; $K_1 \not\subseteq K_2$, $K_2 \not\subseteq K_1$.

Then, $K_1 \cap K_2 = \emptyset$.

Proof. Let $K_1 = E_1 \cup G$, $K_2 = E_2 \cup G$, where E_1 and E_2 are nonvoid disjoint subsets of nodes (Fig. 1). We shall proof that $G = \emptyset$. K_1 is a SMS, then $w(K_1) \leq w(E_1)$, that is:

$$\alpha_G + \alpha_{E_2 G} + \beta_{E_2 G} \leq \beta_{E_1 G} \tag{1}$$

where the equal sign holds for $G = \emptyset$ only.

K_2 is a MS, then $w(K_2) \leq w(E_2)$, that is:

$$\alpha_G + \alpha_{E_1 G} + \beta_{E_1 G} \leq \beta_{E_2 G} . \tag{2}$$

Comparing (1) and (2) it results:

$$\alpha_G = \alpha_{E_1 G} = \alpha_{E_2 G} = 0 \tag{3.a}$$

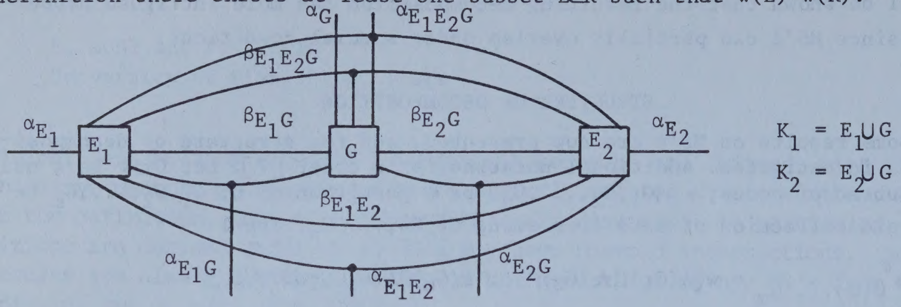
$$\beta_{E_1 G} = \beta_{E_2 G} \tag{3.b}$$

which implies $w(K_1) = w(E_1)$, that is $G = \emptyset$. □

Theorem 1 directly derives from a theorem of [2], which states that SMS's cannot partially overlap. Theorem 1 states that also SMS's and MS's cannot partially overlap; that is, the decomposition in MS's introduces, over the decomposition in SMS's, an additional collection of sets, each of which totally contains SMS's, and is totally contained in SMS's.

Relations between two MS's.

The intersection of two MS's is now studied. Let K be an MS non SMS; by definition there exists at least one subset $E \subset K$ such that $w(E)=w(K)$.



$$w(K_1)=\alpha_{E_1}+\alpha_G+\alpha_{E_1E_2}+\alpha_{E_1G}+\alpha_{E_2G}+\alpha_{E_1E_2G}+\beta_{E_1E_2}+\beta_{E_2G}+\beta_{E_1E_2G}$$

$$w(K_2)=\alpha_{E_2}+\alpha_G+\alpha_{E_1E_2}+\alpha_{E_1G}+\alpha_{E_2G}+\alpha_{E_1E_2G}+\beta_{E_1E_2}+\beta_{E_1G}+\beta_{E_1E_2G}$$

$$w(E_1)=\alpha_{E_1}+\alpha_{E_1E_2}+\alpha_{E_1G}+\alpha_{E_1E_2G}+\beta_{E_1E_2}+\beta_{E_1G}+\beta_{E_1E_2G}$$

$$w(E_2)=\alpha_{E_2}+\alpha_{E_1E_2}+\alpha_{E_2G}+\alpha_{E_1E_2G}+\beta_{E_1E_2}+\beta_{E_2G}+\beta_{E_1E_2G}$$

Fig. 1. Illustration of Theorems 1 and 2.

Proposition 1. Let K be a MS; $E \subset K$ be a proper subset of K; $w(E)=w(K)$. Then, E is a MS.

Proof. Any subset M of E is also a subset of K: then, $w(M) \geqslant w(K)$. It follows $w(M) \geqslant w(E)$, that is E is a MS. □

Theorem 2. Let K_1 and K_2 be MS's; $K_1 \not\subseteq K_2$, $K_2 \not\subseteq K_1$; $K_1 \cap K_2 \neq \emptyset$.
Then $K_1-K_1 \cap K_2$ and $K_2-K_1 \cap K_2$ are MS's, with $w(K_1-K_1 \cap K_2)=w(K_1)$ and $w(K_2-K_1 \cap K_2)=w(K_2)$.

Proof. Let $K_1=E_1 \cup G$, $K_2=E_2 \cup G$, where E_1, E_2, G are nonvoid disjoint subsets (Fig. 1). It is $K_1-K_1 \cap K_2=E_1$, $K_2-K_1 \cap K_2=E_2$, $K_1 \cap K_2=G$. Since K_1 and K_2 are MS's, then $w(K_1) \leqslant w(E_1)$ and $w(K_2) \leqslant w(E_2)$: as a consequence, relations (1), (2) and (3) still hold (see proof of Theorem 1). From (1) and (3) it follows $w(E_1)=w(K_1)$, that is $w(K_1-K_1 \cap K_2)=w(K_1)$. Hence, $K_1-K_1 \cap K_2$ is a MS, by proposition 1.
From (2) and (3) it follows $w(E_2)=w(K_2)$, that is $w(K_2-K_1 \cap K_2)=w(K_2)$. Hence, $K_2-K_1 \cap K_2$ is a MS by proposition 1. □

Corollary 1. Let K_1 and K_2 be MS's; $K_1 \not\subseteq K_2$, $K_2 \not\subseteq K_1$; $K_1 \cap K_2 \neq \emptyset$.
Then, $w(K_1 \cap K_2) \geqslant w(K_1)$, $w(K_2)$.

The proof immediately follows from definition 1.

Corollary 2. Let K_1 and K_2 be MS's; $K_1 \not\subseteq K_2$, $K_2 \not\subseteq K_1$; $K_1 \cap K_2 \neq \emptyset$.
And let $\pi = \{K_1-K_1 \cap K_2, K_1 \cap K_2, K_2-K_1 \cap K_2\}$. Then:

$$\alpha^{\pi}_{K_1 \cap K_2} = 0; \qquad (4.a)$$

$$\alpha^{\pi}_{K_1-K_1\cap K_2,K_1\cap K_2} = 0\,; \tag{4.b}$$

$$\alpha^{\pi}_{K_2-K_1\cap K_2,K_1\cap K_2} = 0\,; \tag{4.c}$$

$$\beta^{\pi}_{K_1-K_1\cap K_2,K_1\cap K_2} = \beta^{\pi}_{K_2-K_1\cap K_2,K_1\cap K_2}\,. \tag{4.d}$$

The proof immediately follows from relations (3.a), (3.b) which are respectively equivalent to relations (4.a,b,c) and (4.d).

Theorem 2, and corollaries 1 and 2, state that two MS's can partially overlap, if special conditions on the sets resulting from the intersection are met. Consider for example the portion of weighted hypergraph in Fig. 2. As it can be directly verified, K_1 and K_2 are MS's with $w(K_1)=6$ and $w(K_2)=5$. Then, E_1 and E_2 are MS's, with $w(E_1)=w(K_1)$ and $w(E_2)=w(K_2)$ (theorem 2). $w(G)=6 \geqslant w(K_1)$, $w(K_2)$ (corollary 1). $\alpha_G=\alpha_{E_1G}=\alpha_{E_2G}=0$, $\beta_{E_1G}=\beta_{E_2G}=3$ (corollary 2).

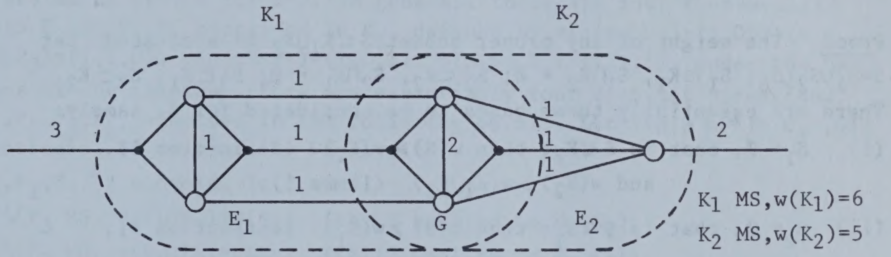


Fig. 2. Sample hypergraph to illustrate the relation of two MS's.

Two theorems will now be presented, leading to the construction of a new MS from two given MS's. Preliminary results are the following.

Proposition 2. Let K be a MS; S be a set of nodes; $K \not\subseteq S$, $S \not\subseteq K$; $K\cap S \neq \emptyset$. Then $w(S) \geqslant w(S-K\cap S)$.

Proof. Let $K=E_1\cup G$, $S=E_2\cup G$ (i.e.: $S-K\cap S=E_2$). It results (see Fig. 1 , where K_1 and K_2 play the role of K and S, respectively):

$$w(S)-w(E_2)=\alpha_G+\alpha_{E_1G}+\beta_{E_1G}-\beta_{E_2G}\,. \tag{5.a}$$

By definition, $w(K) \leqslant w(E_1)$; that is:

$$w(K)-w(E_1)=\alpha_G+\alpha_{E_2G}+\beta_{E_2G}-\beta_{E_1G} \leqslant 0\,. \tag{5.b}$$

From relation (5.b) it follows $\beta_{E_1G} \geqslant \beta_{E_2G}$. Hence, $w(S) \geqslant w(E_2)$ by relation (5.a). □

Lemma 1. Let K_1 and K_2 be MS's; $K_1 \not\subseteq K_2$, $K_2 \not\subseteq K_1$; $K_1\cap K_2 \neq \emptyset$. Then, $w(K_1\cup K_2) \leqslant w(K_1), w(K_2)$.

Proof. Let $K_1=E_1\cup G$, $K_2=E_2\cup G$ (see Fig.1). By corollary 2, it results:

$$w(K_1\cup K_2)=\alpha_{E_1}+\alpha_{E_2}+\alpha_{E_1E_2}+\alpha_{E_1E_2G}; \tag{6.a}$$

$$w(K_1)=\alpha_{E_1}+\alpha_{E_1E_2}+\alpha_{E_1E_2G}+\beta_{E_1E_2}+\beta_{E_2G}+\beta_{E_1E_2G}; \tag{6.b}$$

$$w(K_2)=\alpha_{E_2}+\alpha_{E_1E_2}+\alpha_{E_1E_2G}+\beta_{E_1E_2}+\beta_{E_1G}+\beta_{E_1E_2G}; \qquad (6.c)$$

$$w(G) =\alpha_{E_1E_2G}+\beta_{E_1G}+\beta_{E_2G}+\beta_{E_1E_2G} \cdot \qquad (6.d)$$

From relations (6.c), (6.d) and corollary 1 it follows:

$$w(G)-w(K_2)=\beta_{E_2G}-\alpha_{E_2}-\alpha_{E_1E_1}-\beta_{E_1E_2}\geqslant 0 \text{ , that is } \beta_{E_2G}-\alpha_{E_2}\geqslant 0.$$

From relations (6.a), (6.b), and the relation above, it follows:

$$w(K_1)-w(K_1\cup K_2)=\beta_{E_2G}+\beta_{E_1E_2}+\beta_{E_1E_2G}-\alpha_{E_2}\geqslant 0, \text{ that is } w(K_1\cup K_2)\leqslant w(K_1).$$

Similarly, it results $w(K_1\cup K_2)\leqslant w(K_2)$. □

An important result is now proven, namely the union of two MS's partially overlapping is a MS.

Theorem 3. Let K_1 and K_2 be MS's; $K_1\not\subseteq K_2$, $K_2\not\subseteq K_1$; $K_1\cap K_2 \neq \emptyset$.
Then $K_1\cup K_2$ is a MS.

Proof. The weight of any proper subset $S\subset K_1\cup K_2$ is evaluated. Let $S=S_1\cup S_2\cup S_3$; $S_1\subset K_1$, $S_1\cap K_2 = \emptyset$; $S_2\subset K_2$, $S_2\cap K_1 = \emptyset$; $S_3\subset K_1$, $S_3\subset K_2$.
There are essentially three cases to be considered for S, namely:

(i) $S_1= \emptyset$, that is $S\subseteq K_2$; then $w(S)\geqslant w(K_2)$ (definition 1),
and $w(K_2)\geqslant w(K_1\cup K_2)$ (lemma 1);

(ii) $S_2= \emptyset$, that is $S\subseteq K_1$; then $w(S)\geqslant w(K_1)$ (definition 1),
and $w(K_1)\geqslant w(K_1\cup K_2)$ (lemma 1);

(iii) $S_1\neq \emptyset$, $S_2\neq \emptyset$, that is $S\not\subseteq K_1$, $S\not\subseteq K_2$; then,
$w(S)\geqslant w(S_2)$ (proposition 2 applied to K_1 and S);
$w(S_2)\geqslant w(K_2)$ (definition 1); and $w(K_2)\geqslant w(K_1\cup K_2)$ (lemma 1).

Then, in any case it results $w(S)\geqslant w(K_1\cup K_2)$, that is $K_1\cup K_2$ is a MS. □

In the example of Fig. 2 it can be verified that $w(K_1\cup K_2)\prec w(K_1)$, $w(K_1\cup K_2)=w(K_2)$ (lemma 1), and $K_1\cup K_2$ is a MS (theorem 3).

A mechanism for constructing a new MS from two disjoint MS's is stated in the following theorem.

Theorem 4. Let K_1 and K_2 be MS's; $K_1\cap K_2 = \emptyset$.
If $w(K_1\cup K_2)\leqslant w(K_1), w(K_2)$, then $K_1\cup K_2$ is a MS.

Proof. Let $S=S_1\cup S_2$ be a proper subset of $K_1\cup K_2$, $S_1\subseteq K_1$, $S_2\subseteq K_2$.
One of the three cases must hold:

(i) $S_1= \emptyset$, that is $S\subseteq K_2$; then, $w(S)\geqslant w(K_2)$ (definition 1);

(ii) $S_2= \emptyset$, that is $S\subseteq K_1$; then, $w(S)\geqslant w(K_1)$ (definition 1);

(iii) $S_1\neq \emptyset$, $S_2\neq \emptyset$, that is $S\not\subseteq K_1$, $S\not\subseteq K_2$; then, $w(S)\geqslant w(S_2)$ (proposition 2 applied to K_1 and S); $w(S_2)\geqslant w(K_2)$ (definition 1).

Then, in any case it results $w(S)\geqslant w(K_1)$ or $w(S)\geqslant w(K_2)$. Under the hypothesis that $w(K_1\cup K_2)\leqslant w(K_1), w(K_2)$, it results $w(S)\geqslant w(K_1\cup K_2)$, that is $K_1\cup K_2$ is a MS. □

Relations among n MS's.

More than two MS's may have nonvoid intersections, under very restrictive conditions. To investigate this case we note that, for a collection of MS's $\gamma=\{K_1,K_2,\dots,K_n\}$, every node $p\in\bigcup_{i=1}^{n} K_i$ may be characterized by a n-dimensional binary membership vector $v(p) = (x_1,\dots x_n)$, where $x_i = 0$ if $p\notin K_i$, $x_i = 1$ if $p\in K_i$, $i\in\{1,\dots,n\}$. For any given membership vector v^*, the set $P=\{p\,|\,v(p)=v^*\}$ is a (possibly void) fundamental product of γ.

Theorem 5. Let $\gamma=\{K_1,\dots,K_n\}$ be a collection of MS's, $n>2$; $K_i\not\subseteq\bigcup_{j\in J} K_j$, $i\in\{1,2,\dots,n\}$, $J=\{1,\dots,i-1,i+1,\dots,n\}$; $K_1\cap K_i\neq\emptyset$, $i\in\{2,\dots,n\}$. And let $P_1,P_2,\dots,P_k$ be the nonvoid fundamental products of γ, contained in K_1. Then $P_1,P_2,\dots,P_k$ are MS's, with $w(P_1)=w(P_2)=\dots=w(P_k)=w(K_1)$.

Proof. By mathematical induction.

(1) The theorem is proven for $n=3$. In general, there are four fundamental products P_1,P_2,P_3,P_4 contained in K_1, defined by $v(p\,|\,p\in P_1)=(1,0,0)$, $v(p\,|\,p\in P_2)=(1,1,0)$, $v(p\,|\,p\in P_3)=(1,0,1)$, $v(p\,|\,p\in P_4)=(1,1,1)$. Under the hypotheses of the theorem, there are essentially four distinct situations for P_1,P_2,P_3,P_4, examined in the following points. Let $v(p\,|\,p\in P^*)=(0,1,0)$; by hypothesis, P^* is nonvoid. We have:

(i) P_1,P_2,P_3,P_4 nonvoid. It is:
$P_1\cup P_3$ MS, $w(P_1\cup P_3)=w(K_1)$ (th. 2 applied to K_1,K_2);
$P_1\cup P_2$ MS, $w(P_1\cup P_2)=w(K_1)$ (th. 2 applied to K_1,K_3);
$P_2\cup P^*$ MS (th. 2 applied to K_2,K_3);
$P_1\cup P_3\cup P_4$ MS, $w(P_1\cup P_3\cup P_4)=w(K_1)$ (th. 2 applied to $K_1,P_2\cup P^*$);
$P_3\cup P_4$ MS, $w(P_3\cup P_4)=w(K_1)$ (th. 2 applied to $P_1\cup P_2,P_1\cup P_3\cup P_4$);
P_2 and P_3 MS, $w(P_2)=w(P_3)=w(K_1)$ (th.2 applied to $P_1\cup P_2,P_1\cup P_3$);
P_1 and P_4 MS, $w(P_1)=w(P_4)=w(K_1)$ (th.2 applied to $P_1\cup P_3,P_3\cup P_4$).

(ii) P_1,P_2,P_3 nonvoid; P_4 void. It is:
P_1 MS, $w(P_1)=w(K_1)$ (th.2 applied to $P_1\cup P_2$(see point (i)), K_2);
P_2 and P_3 MS, $w(P_2)=w(P_3)=w(K_1)$ (see point (i)).

(iii) P_1,P_2,P_4 nonvoid; P_3 void. It is:
$P_1\cup P_4$ MS, $w(P_1\cup P_4)=w(K_1)$ (th. 2 applied to $P_2\cup P^*$(see point (i)),K_1);
P_1 MS, $w(P_1)=w(K_1)$ (th. 2 applied to K_1,K_2);
P_2 and P_4 MS's, $w(P_2)=w(P_4)=w(K_1)$ (th.2 applied to $P_1\cup P_2$(see point (i)), $P_1\cup P_4$).

(iv) P_1,P_4 nonvoid; P_2,P_3 void.
$K_1\cup K_2$ MS (th. 3);
$P_1\cup P^*$ MS (th. 2 applied to $K_1\cup K_2$, K_3);
P_1 MS, $w(P_1)=w(K_1)$ (th.2 applied to K_1,K_2);
P_4 MS, $w(P_4)=w(K_1)$ (th.2 applied to $K_1,P_1\cup P^*$).

(2) Assume that the theorem is valid for m, $3\leq m<n$; we prove that it is valid for $m+1$. Let $P_1,\dots,P_h$ be the nonvoid fundamental products of $\{K_1,\dots,K_m\}$, and let $v(p\,|\,p\in P_1)=(1,0,0,\dots,0)$. The fundamental products of $\{K_1,\dots,K_m,K_{m+1}\}$ will be denoted by $P_1^0,P_1^1,\dots,P_h^0,P_h^1$, where $P_i^0\cup P_i^1= P_i$, $v(p\,|\,p\in P_i^0)= (v(p\,|\,p\in P_i),0)$ and $v(p\,|\,p\in P_i^1)=(v(p\,|\,p\in P_i),1)$, $i\in\{1,\dots,h\}$.

(i) For any value of i for which P_i^o (or P_i^1) is void, it is $P_i^1 \equiv P_i$ (or $P_i^o \equiv P_i$), that is P_i^1 (or P_i^o) is a MS with $w(P_i^1)=w(K_1)$ (or $w(P_i^o)=w(K_1)$).

(ii) For any value of $i \neq 1$ for which P_i^1 and P_i^o are nonvoid, it is:

$K_1-K_1\cap K_{m+1}$ MS (th. 2);

$K_1-K_1\cap K_{m+1} \supseteq P_1^o \cup P_i^o$, P_1^o nonvoid by hypothesis;

P_i^o MS, $w(P_i^o)=w(K_1)$ (th. 2 applied to P_i, K_{m+1});

P_i^1 MS, $w(P_i^1)=w(K_1)$ (th. 2 applied to P_i, $K_1-K_1\cap K_{m+1}$).

(iii) If P_1^o and P_1^1 are nonvoid, it is:

P_1^o MS, $w(P_1^o)=w(K_1)$ (see point (ii)).

The collection of nonvoid fundamental products of the MS's $\{K_1,\dots,K_{j-1}, K_{j+1},\dots,K_{m+1}\}$ includes P_1^1; since the theorem is valid for m, then P_1^1 is a MS with $w(P_1^1)=w(K_1)$. □

Further restrictive conditions on the intersection of n MS's are expressed in the following corollaries. For brevity, only the main lines of the proofs will be indicated here.

Corollary 3. Let $\gamma=\{K_1,K_2,K_3\}$ be a collection of MS's; $K_i \not\subseteq \bigcup_{j\in J} K_j$, $i=\{1,2,3\}$, $J=\{1,2,3\} - \{i\}$; $K_1\cap K_i \neq \emptyset$, $i \in \{2,3\}$. And let $\pi=\{P_1,\dots,P_k\}$ be the collection of nonvoid fundamental products of γ contained in K_1.

(1) If $\pi=\{P_1,P_2,P_3\}$, where $v(p|p\in P_1)=(1,0,0)$, $v(p|p\in P_2)=(1,1,0)$, $v(p|p\in P_3)=(1,0,1)$, then all total weights α^π and β^π must be equal to zero, except for:

$$\alpha^\pi_{P_2} = \alpha^\pi_{P_3} = \beta^\pi_{P_1P_2} = \beta^\pi_{P_1P_3};$$

$$\alpha^\pi_{P_1P_2P_3}.$$

(2) For any other collection $\pi=\{P_1,\dots,P_k\}$, all total weights α^π and β^π must be equal to zero, except for $\alpha^\pi_{P_1\dots P_k}$.

The proof can be derived by the application of corollary 2 to all the pairs of MS's contained in K_1, which result from the union of fundamental products. The essentially distinct cases to be considered are the ones listed in the proof of theorem 5, point (1).

Corollary 4. Let $\gamma=\{K_1,K_2,\dots,K_n\}$ be a collection of MS's, $n>3$; $K_i \not\subseteq \bigcup_{j\in J} K_j$, $i=\{1,2,\dots,n\}$, $J=\{1,\dots,i-1,i+1,\dots,n\}$; $K_1\cap K_i \neq \emptyset$, $i\in\{2,\dots,n\}$. And let $\pi=\{P_1,P_2,\dots,P_k\}$ be the collection of nonvoid fundamental products of γ contained in K_1. All total weights α^π and β^π must be equal to zero, except for $\alpha^\pi_{P_1\dots P_k}$.

Corollary 4 is an extension of corollary 3, point (2), and can be proven by mathematical induction from the latter (see proof of theorem 5, point (2)). Note that the extension of corollary 3, point (1), to n MS's (that is, the only nonvoid fundamental products for γ contained in K_1 have vectors with $x_1=1$, and at most one $x_i=1$, $i\in\{2,\dots,n\}$) does not constitute a special case, for $n>3$.

Corollaries 3 and 4 state that, for two collections $\{K_1,\dots,K_n\}$ of MS's and $\{P_1,\dots,P_k\}$ of fundamental products as introduced in theorem 5 - with the exception described in corollary 3, point (1) - no arc leaving any P_i exists, except for arcs common to all P_i's. This condition is quite restrictive in any hypergraph. In particular, in a graph all α's have one index and all β's have two indexes, therefore, no arc leaving any P_i can exist, which implies $w(P_1)=\dots=w(P_k)=w(K_1)=0$. Such a situation is trivial in a non connected graph, and cannot occur if the graph is connected.

The exception relative to corollary 3 point (1) is illustrated in the graph of Figure 3, where: $\alpha_{P_2} = \beta_{P_2P_1} = \beta_{P_1P_3} = \alpha_{P_3} = 3$; $\alpha_{P_1} = 0$.

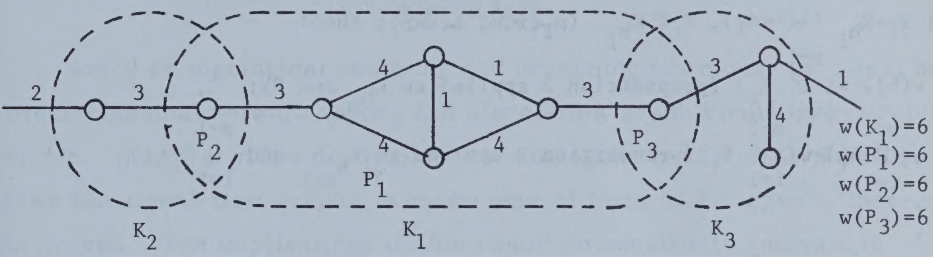


Fig. 3. Sample graph to illustrate corollary 3 point (1).

An extension of theorem 4 can be given, leading to the construction of a new MS from the union of n disjoint MS's. In this perspective, a proposition will be proven first.

Proposition 3. Let $K_1,K_2,\dots,K_m$ be disjoint MS's; S be a set of nodes; $K_1\not\subseteq S$, $S\not\subseteq K_1$, $K_1\cap S\neq\emptyset;\dots;K_m\not\subseteq S$, $S\not\subseteq K_m$, $K_m\cap S\neq\emptyset$; $S-(K_1\cup K_2\cup\dots\cup K_m)\cap S\neq\emptyset$. Then $w(S)\geqslant w(S-(K_1\cup K_2\cup\dots\cup K_m)\cap S)$.

Proof. Let $S_i=S\cap K_i$ $(1\leqslant i\leqslant m)$, $S_{m+1}=S-(K_1\cup K_2\cup\dots\cup K_m)\cap S$; $L_i=\bigcup_{j=1}^{i} S_j\cup S_{m+1}$ $(1\leqslant i\leqslant m)$ (i.e.: $L_m=S$). By mathematical induction:

(1) $w(L_1)\geqslant w(S_{m+1})$ (Proposition 2);

(2) Assume that, for a given $i\leqslant m$, it is $w(L_i)\geqslant w(S_{m+1})$. Since $w(L_{i+1})\geqslant w(L_i)$ (Proposition 2 applied to the MS K_{i+1} and to the set L_{i+1}), it results: $w(L_{i+1})\geqslant w(S_{m+1})$.

The requirements for induction are then met, that is $w(L_m)\geqslant w(S_{m+1})$. □

Theorem 6. Let $K_1,K_2,\dots,K_n$ be MS's, $K_i\cap K_j=\emptyset$, $1\leqslant i<j\leqslant n$. If:

(1) $w(\bigcup_{i=1}^{n} K_i)\leqslant w(K_j)$, $1\leqslant j\leqslant n$;

(2) no subcollection $K_{h_1},K_{h_2},\dots,K_{h_r}$, $1<r<n$, exists, such that $w(\bigcup_{i=1}^{r} K_{h_i})<w(K_j)$, $1\leqslant j\leqslant n$;

then $\bigcup_{i=1}^{n} K_i$ is a MS.

Proof. Let $S=\bigcup_{i=1}^{m} S_i$ $(1\leq m\leq n)$, $S_i\subseteq K_{h_i}$, be a proper subset of $\bigcup_{i=1}^{n} K_i$. One of the following cases must hold:

(i) $S=S_1$ (i.e.: m=1); then $w(S)\geq w(K_{h_1})\geq w(\bigcup_{i=1}^{n} K_i)$ (def.1 and hypot. (1));

(ii) $S_i=K_{h_i}$ $(1\leq i\leq m,\ 1<m\leq n)$; then

$w(S)\geq w(K_j)\geq w(\bigcup_{i=1}^{n} K_i)$ $(1\leq j\leq n)$ (hypotheses (2) and (1));

(iii) $S_i\subset K_{h_i}$ $(1\leq i\leq m,\ 1<m\leq n)$; then $w(S)\geq w(K_{h_m})$ (Proposition 3 applied to $K_{h_1},\ldots,K_{h_{m-1}}$ and S), and

$w(K_{h_m})\geq w(\bigcup_{i=1}^{n} K_i)$ (hypothesis (1));

(iv) $S_i=K_{h_i}$ $(1\leq i\leq m_1)$, $S_i\subset K_{h_i}$ $(m_1<i\leq m,\ 1<m\leq n)$; then:

$w(S)\geq w(\bigcup_{i=1}^{m-1} S_i)$ (proposition 2 applied to K_{h_m} and S);

$w(\bigcup_{i=1}^{m-1} S_i)\geq w(\bigcup_{i=1}^{m-2} S_i)$(proposition 2 applied to $K_{h_{m-1}}$ and $\bigcup_{i=1}^{m-1} S_i$);

.

.

.

$w(\bigcup_{i=1}^{m_1+1} S_i)\geq w(\bigcup_{i=1}^{m_1} S_i)$(prop.2 applied to $K_{h_{m+1}}$ and $\bigcup_{i=1}^{m_1+1} S_i$);

$w(\bigcup_{i=1}^{m_1} S_i)\geq w(K_j)\geq w(\bigcup_{i=1}^{n} K_i)$ $(1\leq j\leq n)$(hypotheses (2) and (1)).

Then, in any case, $w(S)\geq w(\bigcup_{i=1}^{n} K_i)$; that is, $\bigcup_{i=1}^{n} K_i$ is a MS. □

CONCLUSION

In this paper the structure of decomposition of a hypergraph is investigated, under a particular definition of minimal connected subsets of nodes. Such subsets (called MS's) include the sets (SMS's)of a classical cathegory, which induce a tree partition on the set of nodes. However, the decomposition in MS's is much more complex, since MS's can partially overlap under special conditions.

An efficient (i.e., polinomial bounded) algorithm for finding all SMS's is known [3]. It is hoped that some theorems given here will be useful tools to construct an efficient algorithm for the decomposition in MS's.

REFERENCES

[1] C. Berge. Graphes et hypergraphes. Dunod, Paris 1970.

[2] F. Luccio, M. Sami. On the decomposition of networks in minimally interconnected subnetworks. IEEE Transactions on Circuits Theory. CT-16, 184-188 (May 1969).

[3] E.L. Lawler. Cutset and partitions of hypergraphs. Memo No. ERL-M340, Electronic Research Laboratory, University of California, Berkeley, May 1972. Under publication in Networks.

A COMPARATIVE STUDY OF TOPOLOGICAL ANALYSIS METHODS FOR CIRCUITS AND SYSTEMS

A. Y. LEE
Department of Electrical Engineering
Michigan Technological University
Houghton, Michigan 49931

H. D'ANGELO
Department of Mathematics
Memphis State University
Memphis, Tennessee 38111

ABSTRACT

Based on topological summing and branching matrices, parallels between classical network theory and signal-flow graph theory are carefully drawn. Starting with a development of analogous forms of Kirchhoff's laws for signal-flow graphs, a most general form of Tellegen's Theorem is proved. The implications of this result on sensitivity analysis is examined.

INTRODUCTION

It has been recently established that fundamental results of classical network theory can be extended to any system which can be characterized by a signal-flow graph [1, 2, 3] . Although it has been more or less intuitive that such extensions of network theory to signal-flow graphs hinge on the topological similarities of classical networks and signal-flow graphs, a clear parallel development has not yet been given. It appears that the differing topological constraints of signal-flow graphs and classical networks (Kirchhoff's laws) have obscured underlying analogies. Perhaps most importantly, the absence of such analogies, evidenced by differing developments of corresponding results (e. g., Tellegen's Theorem [1, 2]) seems to belie some lack of understanding of the essence of topological structures required to obtain the desired results.

In this paper, the parallels between classical network theory and signal-flow graph theory are carefully drawn. These parallels are based on the topological matrices, summing and branching matrices, which have been defined for signal-flow graphs [4] . These matrices allow analogous forms of Kirchhoff's laws to be written for signal-flow graphs. These formulations lead to a most general form of Tellegen's Theorem for signal-flow graphs.

PRELIMINARY

A signal-flow graph of a system is a collection of nodes and oriented branches. When drawing the branch, an arrow is used to indicate the orientation of the branch. An input node is one which has one and only one outgoing branch, an input branch. Associated with an input branch is an input variable, u. An output node is one which has one and only one incoming branch, an output branch. Associated with an output branch is an output variable, y. All other nodes besides the input and output nodes have a degree of at least two and are called n-nodes. For each of the n-nodes, a variable n is introduced. All other branches besides the input and output branches are called f-branches. There are two variables related by a function associated with each of the f-branches. The variable at the tail of the branch is referred to as the initial variable, x, and that at the head of the branch is referred to as the final variable, v. The relationship between the initial and final variables for branch i is

$$v_i = f_i(x_i). \tag{1}$$

The summing matrix and branching matrix of an m-input, r-output signal-flow graph are defined next. These matrices completely describe the topological structure of the signal-flow graph. Let the signal-flow graph consist of p+m+r nodes (m input nodes, r output nodes, and p n-nodes), and q+m+r branches (m input branches, r output branches, and q f-branches). The order of labeling the nodes and branches is immaterial except that the following method of labeling simplifies the explanation of the formulation of the summing and branching matrices. For the branches, first label the f-branches from 1 to q; next, label the input branches from q+1 to q+m corresponding to the input variables $u_1, u_2, \ldots, u_m$, respectively; finally, label the output branches from q+m+1 to q+m+r corresponding to the output variables $y_1, y_2, \ldots, y_r$, respectively. For the nodes, first label the n-nodes from 1 to p, next the input nodes, and finally the output nodes.

The summing matrix, $S = [S_1 \mid S_2]$ is a p x (q+m) matrix. The partition S_1, a p x q matrix, is defined by

$$s_{ij} = \begin{cases} 1 & \text{if node i is the final node of f-branch j for } j = 1, \ldots, q \\ 0 & \text{otherwise} \end{cases} \qquad (2a)$$

The partition S_2, a p x m matrix, is defined by

$$s_{ij} = \begin{cases} 1 & \text{if node i is the final node of input branch j for } j = q+1, \ldots, q+m^{*} \\ 0 & \text{otherwise} \end{cases} \qquad (2b)$$

Thus

$$S_1 = \begin{bmatrix} s_{11} \cdots & s_{1q} \\ \vdots & \\ s_{p1} \cdots & s_{pq} \end{bmatrix} \qquad (3a)$$

and

$$S_2 = \begin{bmatrix} s_{1\ q+1} \cdots & s_{1\ q+m} \\ \vdots & \vdots \\ s_{p\ q+1} \cdots & s_{p\ q+m} \end{bmatrix} \qquad (3b)$$

The branching matrix, $B = \begin{bmatrix} B_1 \\ \hline B_2 \end{bmatrix}$ is a (q+r) x p matrix. The partition B_1, a q x p matrix, is defined by

$$b_{ij} = \begin{cases} 1 & \text{if node j is the initial node of f-branch i for } i=1, \ldots, q \\ 0 & \text{otherwise} \end{cases} \qquad (4a)$$

The partition B_2, a r x p matrix, is defined by

$$b_{ij} = \begin{cases} 1 & \text{if node j is the initial node of the output branch m+i for } i=q+1, \ldots, q+r^{**} \\ 0 & \text{otherwise} \end{cases} \qquad (4b)$$

---

*This is for the case that the input branches are numbered before the output branches. If the opposite is true, then $s_{ij} = 1$ if node i is the final node of the input branch r + j for $j = q+1, \ldots, q+m$.

**This is the case that the input branches are numbered before the output branches. If the opposite is true, then $b_{ij} = 1$ if node j is the initial node of the output branch i for $i=q+1, \ldots, q+r$.

Thus

$$B_1 = \begin{bmatrix} b_{11} \cdots & b_{1p} \\ \vdots & \vdots \\ b_{q1} \cdots & b_{qp} \end{bmatrix} \tag{5a}$$

and

$$B_2 = \begin{bmatrix} b_{q+1\ 1} \cdots & b_{q+1\ p} \\ \vdots & \vdots \\ b_{q+r\ 1} \cdots & b_{q+r\ p} \end{bmatrix} \tag{5b}$$

As an example of how the summing and branching matrices are formed, consider the 2-input, 3-output signal-flow graph of Figure 1. The summing matrix and branching matrix for this graph are respectively

$$S = \left[\begin{array}{ccccccccc:cc}
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot
\end{array}\right]$$

$$B = \left[\begin{array}{ccccccc}
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\
\hdashline
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & , \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1
\end{array}\right]$$

where the $\cdot$ indicates a zero entry. Notice that every column of the S matrix contains exactly one nonzero entry and that every row of the B matrix contains exactly one nonzero entry.

PARALLEL DEVELOPMENTS

1. Kirchhoff's Laws

Recall that the constraints for classical networks are Kirchhoff's laws:

$$\text{(a)}\quad Ci = 0 \quad \text{and} \quad \text{(b)}\ Fv = 0 \tag{6}$$

where C and F are the fundamental seg matrix and the fundamental circuit matrix respectively and i and v are the branch current and branch voltage vectors respectively [5]. These equations can be rewritten as:

$$\text{(a)} \quad [I \;\; C_c]\begin{bmatrix} i_t \\ i_c \end{bmatrix} = 0 \quad \text{and} \quad \text{(b)} \quad [F_t \;\; I]\begin{bmatrix} v_t \\ v_c \end{bmatrix} = 0 \tag{7}$$

where the t and c subscripts denote the tree and cotree portions of the network graph respectively and I is the identity matrix.

Corresponding results are obtained for signal-flow graphs. Let $n = [n_1 \; n_2 \ldots n_p]^T$, $x = [x_1 \; x_2 \ldots x_q]^T$, $v = [v_1 \; v_2 \ldots v_q]^T$, $u = [u_1 \; u_2 \ldots u_m]^T$, and $y = [y_1 \; y_2 \ldots y_r]^T$, where T indicates the transpose of a matrix. The vector n is defined by the following equation:

$$n = [S_1 \;\; S_2]\begin{bmatrix} v \\ u \end{bmatrix} \tag{8}$$

x and y can be expressed as a linear function of n

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} n \tag{9}$$

and v and x are related through the branch relations

$$v_i = f_i(x_i), \qquad i = 1, \ldots, q \tag{10}$$

Defining $w = \begin{bmatrix} v \\ u \end{bmatrix}$ and $z = \begin{bmatrix} x \\ y \end{bmatrix}$, equations 8 and 9 become:

$$\text{(a)} \quad n = Sw \quad \text{and} \quad \text{(b)} \quad z = Bn. \tag{11}$$

These can be rearranged to a form corresponding to Kirchhoff's laws for classical networks:

$$[I \quad S]\begin{bmatrix} -n \\ w \end{bmatrix} = 0 \quad \text{and} \quad [B \quad I]\begin{bmatrix} -n \\ z \end{bmatrix} = 0$$

2. Tellegen's Theorem

Similarly, recall that Tellegen's Theorem [6] states that for any two networks N and $\hat{N}$ having identical topologies, the following relations hold:

$$i^T\hat{v} = 0, \quad \hat{i}^T v = 0$$

Corresponding results are obtained for signal-flow graphs. Consider the two p+m+r node, q+m+r branch signal-flow graphs N and $\hat{N}$ represented in Figure 2. Also note that for $\hat{N}$ much of the defined nomenclature is reversed. Namely $\hat{N}$ has r input nodes and m output nodes, $\hat{y}_i$ is an input variable and $\hat{u}_i$ is an output variable, and $\hat{v}_i$ is an initial f-branch variable and $\hat{x}_i$ is a final f-branch variable. N and $\hat{N}$ are otherwise unrelated.

However, if N and $\hat{N}$ have identical topologies with flows reversed, then $\hat{N}$ is said to be the reverse graph of N. In this case, the summing and branching matrices of $\hat{N}$, $\hat{S}$ and $\hat{B}$ respectively, are simply related to the summing and branching matrices of N: $\hat{S} = B^T$ and $\hat{B} = S^T$.

Defining $\hat{w} = \begin{bmatrix} \hat{v} \\ \hat{u} \end{bmatrix}$ and $\hat{z} = \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix}$, one obtains

$$\text{(a)}\ \hat{n} = \hat{S}\,\hat{z} \quad \text{and} \quad \text{(b)}\ \hat{w} = \hat{B}\,\hat{n}. \tag{12}$$

Based on these results a theorem corresponding to Tellegen's Theorem is proved:

Theorem 1: If $\hat{N}$ is reversed graph of N, then $[\hat{n}^T\ \hat{z}^T] \begin{bmatrix} -n \\ z \end{bmatrix} = 0$ and $[\hat{n}^T\ \hat{w}^T] \begin{bmatrix} -n \\ w \end{bmatrix} = 0$.

Proof: Since $\hat{S} = B^T$ and $\hat{B} = S^T$, $\hat{n}^T n = \hat{n}^T S w = (S^T \hat{n})^T w = (\hat{B}\hat{n})^T w = w^T w$. Also, $\hat{n}^T n = \hat{z}^T \hat{S}^T n = \hat{z}^T (n^T \hat{S})^T = \hat{z}^T (n^T B^T)^T = \hat{z}^T B n = \hat{z}^T z$. This completes the proof.

3. A Generalized Tellegen's Theorem

A more general form of Theorem 1 not requiring N and $\hat{N}$ to be reverse graphs of each other is obtained:

Theorem 2: For any two signal-flow graphs N and $\hat{N}$ having the same number of nodes and branches

$$\begin{bmatrix} \hat{n} \\ \hat{w} \end{bmatrix}^T \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} -n \\ w \end{bmatrix} = 0 \tag{13}$$

and

$$\begin{bmatrix} \hat{n} \\ \hat{z} \end{bmatrix}^T \begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} -n \\ z \end{bmatrix} = 0 \tag{14}$$

where A_{11} and H_{11} are arbitrary p x p matrices, and where

$$A_{22} = \hat{B}^{+T} A_{11} S$$

$$H_{22} = \hat{S}^T H_{11} B^+$$

where the + indicates the generalized inverse.

Proof: From equations (11a) and (12b), $\hat{n}^T A_{11} n = \hat{w}^T \hat{B}^{+T} A_{11} S w$. Defining $A_{22} = \hat{B}^{+T} A_{11} S$ results in equation 13. Similarly from equations (11b) and (12a), $\hat{n}^T H_{11} n = \hat{z}^T \hat{S}^T H_{11} B^+ z$. Defining $H_{22} = \hat{S}^T H_{11} B^+$ results in equation 14.

Theorem 2 suggests that a generalized form of Tellegen's Theorem also exists for classical networks:

Theorem 3: For any two networks N and $\hat{N}$ having the same number of nodes and branches

$$\begin{bmatrix} i_t \\ i_c \end{bmatrix}^T \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} \hat{v}_t \\ \hat{v}_c \end{bmatrix} = 0$$

and

$$\begin{bmatrix} \hat{i}_t \\ \hat{i}_c \end{bmatrix}^T \begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} v_t \\ v_c \end{bmatrix} = 0$$

where A_{11} and H_{11} are arbitrary square matrices whose dimensions are the same as the rank of the network graphs, and where

$$A_{22} = -C_c{}^T A_{11} \hat{F}_t{}^+$$

and $$H_{22} = -\hat{C}_c{}^T H_{11} F_t{}^+$$

The proof is similar to the proof of Theorem 2 and follows immediately from Kirchhoff's laws (equations 7a and 7b).

Note that for any network $-F_t{}^T = C_c$ and thus $A_{22} = C_c{}^T A_{11} \hat{C}_c{}^{+T}$ and $H_{22} = \hat{C}_c{}^T H_{11} C_c{}^{+T}$.

Note, if $A_{22} = I$, $H_{22} = I$, and networks N and $\hat{N}$ have the same topology, then Theorem 3 reduces to the well-known form of Tellegen's Theorem. Similarly, for signal-flow graphs, if $A_{22} = I$, $H_{22} = I$, and signal-flow graphs N and $\hat{N}$ are the reverse graphs of each other, Theorem 2 reduces to Theorem 1.

4. Sensitivity

The application of Tellegen's Theorem to the sensitivity analysis of networks is well known [6, 7]. Using Theorem 1, adjoint signal-flow graphs can be defined and similar sensitivity results can be obtained for signal-flow graphs [3]. The implications of Theorems 2 and 3 on sensitivity analysis are presently being studied and will be discussed in a forthcoming paper.

CONCLUSIONS

Based on topological summing and branching matrices, analogously defined for classical networks and signal-flow graphs, parallel developments of Kirchhoff's Laws and Tellegen's Theorems are advanced. These formulations provide important insights to key links between the mathematical and geometric structures and lead to a most general form of Tellegen's Theorem.

REFERENCES

1 A. Y. Lee, "Topological Properties of the Summing and Branching Matrices of Signal Flow Graphs and the Application to Sensitivity Analysis," Proc. of Sixth Asilomar Conference on Circuits and Systems, November 1972.

2 A Fettweis, "A General Theorem for Signal-Flow Networks, With Applications," ARCH. ELECTR. UBERTR., Vol. 25, Pp. 557-561, December 1971.

3 A. Y. Lee, "Signal Flow Graphs: Computer-Aided System Analysis and Sensitivity Calculations," IEEE Trans. on Circuit Theory, Vol. CT-21, March 1974.

4 A. Y. Lee and H. D'Angelo, "Extensions of Network Topology Concepts to Multivariable Systems," Proc. of the Fifth Annual Southeastern Symposium on System Theory, March 1973.

5 S. Seshu and M. Reed, Linear Graphs and Electrical Networks, Reading, Massachusetts: Addison-Wesley, 1961.

6 D. A. Calahan, Computer Aided Network Design (revised ed.), New York: McGraw-Hill, 1972.

7 S. W. Director and R. A. Rohrer, "A Generalized Adjoint Network and Network Sensitivities," IEEE Trans., Vol. CT-16, Pp. 318-323, August 1969.

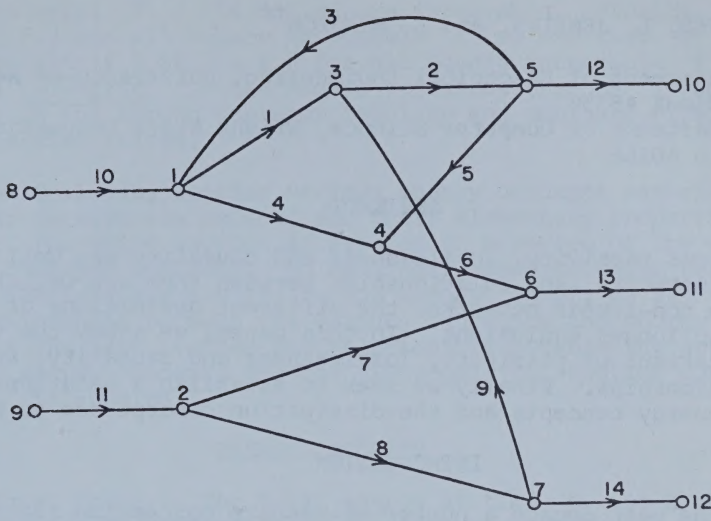

3
3
2
5
12
10
1
5
8
10
1
4
4
6
6
13
11
7
9
9
11
2
8
7
14
12

Figure 1. A 2-input, 3-output signal flow graph

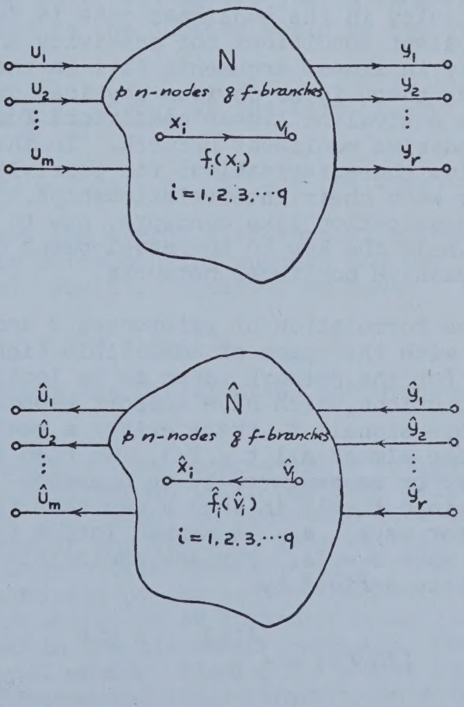

N
u_1
u_2
u_m
p n-nodes q f-branches
x_i
v_i
$f_i(x_i)$
$i = 1, 2, 3, \cdots q$
y_1
y_2
y_r
$\hat{N}$
$\hat{u}_1$
$\hat{u}_2$
$\hat{u}_m$
p n-nodes q f-branches
$\hat{x}_i$
$\hat{v}_i$
$\hat{f}_i(\hat{v}_i)$
$i = 1, 2, 3, \cdots q$
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_r$

Figure 2. Two p+m+r node, q+m+r branch signal-flow graphs N and $\hat{N}$.

PASSIVE NONLINEAR NETWORKS

R. SAEKS[†], L. JENKINS[†], and L. WINSLOW[††]

[†] Department of Electrical Engineering, University of Notre Dame, Indiana 46556

[††] Department of Computer Science, Wright State University, Ohio 40166

ABSTRACT

The terms passivity, losslessness and causality are well defined for linear networks, and relationships between them are well-known. However, in non-linear networks, the different definitions of the terms are no longer equivalent. In this paper, we study the various characterizations of passivity, losslessness and causality, and their inter-relationships. Finally we seek to establish a relationship between our energy concepts and the dissipation concept due to Willems.

INTRODUCTION

Over the past decade a number of results concerning the relationship between causality and various energy concepts such as losslessness, passivity and stability have been obtained for linear networks[1,2,3,10]. For nonlinear networks, however, similar relationships have been exhibited only in a number of restricted circumstances. In particular, the lossless case[4] and the weakly additive case[4,5,6] have yielded relationships analogous to those exhibited by linear networks. The difficulty in the nonlinear case is due to the fact that the various equivalent conditions for passivity and causality which are used routinely in linear arguments fail in the case of nonlinear networks. Indeed, there is no obvious physical criterion by which to choose one of the equivalent linear conditions for passivity as "the" definition of a passive nonlinear network. In the present paper the various alternative characterizations for passivity and causality are surveyed together with their interrelationships. Included among these are a number of dissipation like concepts, new to the present paper, which we believe hold the key to the development of a definitive characterization of passive nonlinear networks.

Following the formulation of references 3 and 4 we adopt a scattering formalism with the space of admissible signals (incident and reflected waves) for the network taken to be locally integrable (n-vector valued) functions which have support bounded on the left (i.e. for each admissible signal f there exists a real number T_f such that $f(t) = 0$ for almost all $t < T_f$). We then say that a network is solvable if it may be represented by an operator S mapping this space of signals into itself in such a way that (globally) square integrable incident wave, a, is mapped into a (globally) square integrable reflected wave $b = Sa$. For any admissible signal f we denote by $(f)^t$ its truncate defined by

$$(f)^t(s) = \begin{cases} f(s) & s \le t \\ 0 & s > t \end{cases} \tag{1.1}$$

and for any square integrable f we denote by $\| f \|$ its L_2 norm interable for all f and t. This fact together with a solvability assumption on S assuring that $S(f)^t$ is square integrable. As such,

$\|(f)^t\|$, $\|Sf)^t\|$ and $\|S(f)^t\|$ are well defined though $\| f \|$ or $\| SF \|$ may fail to exist if f is not square integrable. Finally, if $\| Sf \| \leq \| f \|$ for all square integrable f we say that S is *contractive* and if $\| SF \| = f \|$ for all square integrable f we say that it is *isometric*. In the sequel we assume that S is solvable, continuous in the strong operator topology and unbiased (i.e. $S(0)= 0$) unless otherwise stated.

In the following section various energy concepts associated with a nonlinear network are defined and their elementary properties are developed. The third section is devoted to a review of the various causality concepts applicable to nonlinear networks while the fourth and fifth sections are devoted to the exposition of, we believe, all known relationships between the various causality and energy concepts. Finally, in the last section the relationship between the preceding "external" energy concepts and the "internal" dissipation concept due to Willems[7] is studied.

ENERGY CONCEPTS

1. *Total Energy*: The *total energy* at time t, for a network S having input a is defined as

$$W(a,t) = \| (a)^t \|^2 - \|(Sa)^t \|^2 \tag{2.1}$$

Here our assumptions on the signal space assures that the norms in equation (2.1) are finite. Intuitively, total energy is the difference between the energy delivered to the system from the outside world and the energy delivered to the outside world by the system in the time interval $(-\infty,t)$.

The *differential total energy* at the time t for a system S having inputs a and $\underline{a}$ is

$$W(a,\underline{a},t) = \| (a - \underline{a})^t \|^2 - \| (Sa - S\underline{a})^t \|^2 \tag{2.2}$$

i.e. $W(a,\underline{a},t)$ compares the energy differential between two distinct system inputs and their corresponding outputs and may be viewed as the usual nonlinear form of $W(a,t)$. Clearly, we have

$$W(a,\underline{a},t) = W(a - \underline{a}, t) \tag{2.3}$$

in the linear case wherein the differential energy concept is redundant.

2. *Recoverable Energy*: The *unrecoverable energy* associated with a network or system is that energy which has been delivered to the network before time t and cannot be recovered from the network by application of any input in the future. This is essentially an energy dissipation concept though viewed from the point of view of the externally accessible terminals of the network rather than actual internal "loss". As such energy in an infinite delay line, though actually not dissipated in the classical sense, is unrecoverable from the point of the external world. Similarly, energy stored in a capacitor which is then disconnected (for instance, in a switched network) is not dissipated but it is unrecoverable by the outside world. Formally, we define the unrecoverable energy for a network S at time t with input a as

$$D(a,t) = \inf_{\substack{(x)^t=(a)^t \\ x \in L_2}} \left[\| x \|^2 - \|Sx\|^2 \right] \tag{2.4}$$

That is, D(a,t) is the least amount of energy which must of necessity be lost over all time given that the input up till time t is equal to a.

The recoverable energy of S at time t with input a is the natural "compliment" of the unrecoverable energy, i.e. that part of the total energy which can be recovered by proper choice of the part of the input after t, and is defined via

$$F(a,t), = W(a,t) - D(a,t) \tag{2.5}$$

1. Proposition: The recoverable and uncrecoverable enegy in a network satisfy the following conditions:

- i). D(a,t) is a non-decreasing function in t for each a.
- ii). $F(a,\infty) = 0$ for all a.
- iii). $D(a,\infty) = W(a,\infty)$ for all a.
- iv). $D(a,t) \geq 0$ for all a and t if and only if S is contractive.
- v). $D(a,t) = 0$ for all a and t if and only if S is isometric.

Here

$$D(a,\infty) = \lim_{t \to \infty} D(a,t) = \|a\|^2 - \|Sa\|^2 \tag{2.6}$$

where the validity of the second equality follows from the strong continuity of S and the truncation operator. $W(a,\infty)$ and $F(a,\infty)$ are, of course, similarly defined. Note that the monitonicity condition of i) is reminiscent of an entropy condition and, indeed, D(a,t) exhibits many "entropy-like" properties.

3. Relaxation Energy: The relaxation loss for a network S, at time t with input a is the energy lost over all time if the input a is applied up till time t and zero input is applied thereafter. Formally, we have

$$\underline{D}(a,t) = \left[\|(a)^t\|^2 - \|S(a)^t\|^2 \right] \tag{2.7}$$

defining the relaxation loss and

$$\underline{F}(a,t) = W(a,t) - \underline{D}(a,t) \tag{2.8}$$

defining its "complement", the relaxation storage. Of course, we may define $\underline{D}(a,\infty)$ and $\underline{F}(a,\infty)$ by taking limits.

2. Proposition: The relaxation energy functions satisfy:

- i). $\underline{F}(a,\infty) = 0$ for all a.
- ii). $\overline{D}(a,\infty) = W(a,\infty)$ for all a.
- iii). $\overline{D}(a,t) \geq 0$ for all a and t if and only if S is contractive.
- iv). $D(a,t) = 0$ for all a and t if and only if S is isometric.

Finally, the relationship between the relaxation energy and the recoverable energy for a system is given by:

3. Proposition:

i). $D(a,t) \leq \underline{D}(a,t)$ for all a and t.
ii). $F(a,t) \leq \overline{F}(a,t)$ for all a and t.
iii). $\overline{D}(a,t) = \underline{D}(a,t) = 0$ and $F(a,t) = \underline{F}(a,t) = W(a,t)$ for all a and t if S is isometric.

Here the inequality i) follows from the fact that $D(a,t)$ is the infimum over a set of numbers which include $\underline{D}(a,t)$ whereas ii) and iii) are simply algebraic variations on the difinitions and i).

4. Abstract Energy Pairs: Finally the above described concepts can be unified via the formalism of an abstract energy pair, $\overline{D}(a,t)$ and $\overline{F}(a,t)$, which we define to be an arbitrary pair of functions dependent on $(a)^t$ satisfying

$$\overline{D}(a,t) + \overline{F}(a,t) = W(a,t) \tag{2.9}$$

Indeed, it is this abstract energy concept which yields the link between our external energy concepts and Willems' internal energy concepts[7].

CAUSALITY

One of the most fundamental concepts of network theory is that of causality. Intuitively we say a network is causal if the effect (output) does not precede the cause (input), i.e. if the output only depends on the previous and present inputs, and not on the future inputs. This philosophical concept can be formalized mathematically in several different ways all of which are equivalent for linear networks. For nonlinear networks, however, this is not the case and we are forced to consider three different causality concepts; their definition and the relationship between them being the subject of the present section.

The most common formulation of the causality concept is same-input same-output causality (SISO-causality) wherein we require that

$$(Sa)^t = (S\underline{a})^t \tag{3.1}$$

for any two pinuts, a and $\underline{a}$, satisfying

$$(a)^t = (\underline{a})^t \tag{3.2}$$

for some t. This definition implies that an SISO-causal system can not predict the future, for if two inputs are identical before t then the resultant outputs are also identical before t (thus the output before t is independent of the input after t or, equivalently, future inputs cannot effect past outputs).

A second formulation of the causality concept which is non-standard though useful for our applications[4,8] is norm causality (N-causality) wherein we require that the inequality

$$\|(Sa)^t\|^2 \leq \|S(a)^t\|^2 \tag{3.3}$$

be satisfied for all a and t. This concept deals only with the energy in the various inputs and outputs rather than the signals themselves and in some sense says that energy cannot be transmitted backwards in time though it apparently leaves open the possibility of transmitting information backwards in time.[Δ]

Our third and final causality concept is zero-input zero-output causality (ZIZO-causality), which is essentially the special case of SISO-causality wherein one takes $\underline{a} = 0$. Formally we require that

$$(Sa)^t = 0 \tag{3.4}$$

whenever

$$(a)^t = 0 \tag{3.5}$$

for some t., i.e. if the input is zero prior to time t then the resultant output is also zero prior to time t.

The relationships between the various causality concepts were formulated in reference 4.

4. Proposition:

i). If a network is SISO-causal then it is N-causal.
ii). If a network is N-causal then it is ZIZO-causal.
iii). If a network is linear all three causality concepts are equivalent.

Note that ii) is dependent on our assumption that S(0) = 0 though i) is valid without this assumption. Finally, we note that iii) can be generalized to the case of weakly additive networks which are defined by the equality

$$S(a) = S(a)^t + S(a)_t \tag{3.6}$$

where $(a)_t = a - (a)^t$. Although the weakly additive networks constitute only a small subset of the totality of nonlinear networks they include all linear networks and many of the nonlinear networks commonly encountered. Intuitively, a network is weakly additive if it has linear time behavior though it may be nonlinear in other respects and, indeed, weakly additive networks are known to exhibit convolution and frequency domain representations similar to those exhibited by linear networks[5,9] as well as similar causality and passivity characteristics[4,6,8]. Using the weak additivity concept it is shown in references 4 and 6 that:

5. Proposition: If a network is weakly additive all three causality concepts are equivalent.

[Δ] For linear networks this is not the case though examples of N-causal nonlinear networks wherein information may be transmitted backwards in time are known.

PASSIVITY

The various energy and causality concepts which we have defined can all be used to characterize passive nonlinear networks. Indeed, in the linear case they yield equivalent passivity concepts and hence any one approach may be adopted as "the" definition of passivity. In the nonlinear case, however, the various energy and causality concepts yield distinct, though closely related, passivity concepts none of which can be distinguished as "the" correct definition. As such, we present eight separate passivity concepts, together with their interrelationships, without adopting one as the definition. Our main theorem is as follows:

Theorem: Each of the following statements implies all succeeding statements.

i). $W(a,\underline{a},t) \geq 0$ for all a, $\underline{a}$ and t.
ii). S is SISO-causal and Lipschitz contractive.
iii). S is SISO-causal and contractive.
iv). S is N-causal and contractive.
v). $D(a,t) \geq 0$ and $F(a,t) \geq 0$ for all a and t.
vi). $\overline{D}(a,t) \geq 0$ and $\overline{F}(a,t) \geq 0$ for all a and t.
vii). $W(a,t) \geq 0$ for all a and t.
viii). S is ZIZO-causal and contractive.

Moreover, i) and ii) are equivalent, iv and v) are equivalent and all of the conditions are equivalent if S is weakly additive (hence also linear). Note that by Lipschitz contractive we mean the usual nonlinear differential contraction condition that

$$||Sa - Sa|| \leq ||a - \underline{a}|| \tag{4.1}$$

for all a and $\underline{a}$. Finally, we note that vii) is the usual inequality used to define passivity[1,2,3,10] hence all eight conditions are equivalent to the classical passivity concept in the linear case.

Proof: i) ii). The equivalence of these two conditions was verified in reference 11.

ii) iii). Since we assume that $S(0) = 0$ if we take $\underline{a} = 0$ we have

$$||Sa|| = ||Sa - S(0)|| \leq ||a - 0|| = ||a||$$

showing that Lipschitz contractive implies contractive and thus verifying the required implication.

iii) iv). This is simply a statement of condition i) of propotion 4.

iv) v). We already know that S is contractive if and only if $D(a,t) \geq 0$ for all a and t (Proposition 1, part iv), whereas the inequality

$$|| (Sa)^t ||^2 \leq || S(a)^t||^2 \tag{4.2}$$

is clearly equivalent to the equality

$$0 \le \underline{F}(a,t) = W(a,t) - D(a,t) = \|(a)^t\|^2 - \|(Sa)^t\|^2 - [\|(a)^t\|^2 - \|S(a)^t\|^2] = \|S(a)^t\|^2 - \|(Sa)^t\|^2 \tag{4.3}$$

hence norm causality and positive relaxation energy are equivalent as required[Δ].

v) vi). By property ii) of proposition 3

$$F(a,t) \ge \underline{F}(a,t) \tag{4.4}$$

hence $\underline{F} \ge 0$ implies that $F \ge 0$. On the other hand if $\underline{D} \ge 0$ then we have

$$D(a,t) = \underset{\substack{x \in L_2 \\ (a)^t=(x)^t}}{\text{infimum}} [\|x\|^2 - \|S(x)\|^2] = \underset{\substack{x \in L_2 \\ (a)^t=(x)^t}}{\text{infimum}} [D(x,\infty)] \ge 0 \tag{4.5}$$

completing the desired implication.

vi) vii). If $F \ge 0$ and $D \ge 0$ then clearly

$$W(a,t) = F(a,t) + D(a,t) \ge 0 \tag{4.6}$$

vii) viii). This implication was verified in reference 4 and will not be repeated here.

Finally, if the network is weakly additive ZIZO-causality implies SISO causality (Proposition 5) all of the conditions are equivalent.

LOSSLESSNESS

Unlike passive networks the classical equivalent characterizations for linear lossless networks[4] remain valid in the nonlinear case providing one works with norm-causality. In fact, additional equivalent conditions formulated in terms of relaxation energy and recoverable energy are also possible.

Theorem: The following are equivalent.

i). $W(a,t) \ge 0$ and $W(a,\infty) = 0$ for all a and t.
ii). S is norm-causal and isometric.
iii). $\underline{D}(a,t) = 0$ and $\underline{F}(a,t) \ge 0$ for all a and t.
iv). $\overline{D}(a,t) = 0$ and $\overline{F}(a,t) \ge 0$ for all a and t.

Note that i) is the classical definition of losslessness[2,10,11] which, since it remains equivalent to the other reasonable definitions for losslessness may be unambiguously taken as the definition in the nonlinear case.

Proof: The proof is similar to that of the passivity theorem.

[Δ] The relationship between stored energy and causality implied by this result is, in fact, to be expected since $\underline{F}(a,\infty) = 0$ implies that any energy which is stored in finite time must be returned to the outside world at infinity. As such if negative energy were to be stored in finite time (i.e., the network delivers energy to the outside world) the network would have to be assured that energy will be returned by infinity which in turn requires that the network know the future.

INTERNAL ENERGY

All of the energy concepts thus far studied have been external in nature being expressed as relationships between the network inputs and outputs without reference to the internal structure of the network. This is in sharp contrast to the work of Willems[7] wherein dissipative and stored energy are defined in terms of the internal state model of the network. Surprisingly, however, the two approaches are intimately related.

By a state model for a network S we assume that a state space X is given together with a family of transition operators (defined over both finite and semi-infinite intervals) satisfying the "usual axioms for a state model[7,12]". We then say that the network is dissipative[7] if there exists an energy storage function Q mapping the state space to the non-negative real numbers such that whenever the network starts in initial state $X(t_o)$ at time t_o and is brought to final state $X(t_1)$ at time t_1 by the input a then

$$Q(X(t_1)) \leq W(a,t_1,t_o) + Q(X(t_o)) \tag{6.1}$$

where

$$W(a,t_1,t_o) = W(a,t_1) - W(a,t_o) \tag{6.2}$$

Intutively, we assume that the stored energy in a network at time t is entirely determined by its state at that time whence $Q(X(t_1))$ represents the stored energy at time t_1 and $Q(X(t_o))$ represents the energy store in the network at time t_o. $W(a,t_1,t_o)$ represents the energy delivered to the network from the outside world between t_o and t_1. The difference of the quantities on the two sides of 6.1 may be taken as the dissipated energy in the network. Note that Q does not truly represent the physically stored energy in the network. Except for its dependence on state, Q is really just as much an external energy measure as our other concepts in that it is an estimate of the internally stored energy predicated on external observations rather than the actual internal energy. The relationship between Willems' "internal" dissipation concept and the previous concepts may be derived with the aid of the following Lemma.

Lemma A network characterized by a controllable state model is dissipative in the sense of equation (6.1) if and only if there exists an abstract energy pair $\bar{F}(a,t)$ and $\bar{D}(a,t)$ for the network such that

$$\bar{F}(a,t) \geq 0, \quad \bar{D}(a,t) \text{ is non-decreasing in } t.$$

Proof: If the network is dissipative and input a yields state $X(t)$ at time t then we define $\bar{F}$ by the equality

$$\bar{F}(a,t) = Q(X(t)) \tag{6.3}$$

and

$$\bar{D}(a,t) = W(a,t) - \bar{F}(a,t) \tag{6.4}$$

$\bar{F}$ and $\bar{D}$ are dependent only on $(a)^t$ and clearly sum to W hence they form an abstract energy pair for the network. Now since the network is dissipative $Q \geq 0$ implying that $\bar{F} \geq 0$. On the other hand if $t_o \leq t_1$ we have

$$\bar{D}(a,t_o) = W(a,t_o) - \bar{F}(a,t_o) = W(a,t_o) - Q(X(t_o)) \le W(a,t_o) + W(a,t_1,t_o) - Q(X(t_1)) = D(a,t_1) \qquad (6.5)$$

showing that $\bar{D}$ is non-decreasing.

Conversely, if there exists an abstract energy pair, $\bar{F}$ and $\bar{D}$, with $\bar{F}$ positive and $\bar{D}$ non-decreasing we may let

$$Q(X(t)) = \bar{F}(a,t) \qquad (6.6)$$

where a is any input that drives the state of the network to $X(t)$ at time t (which exists if the state model is controllable and is assured to be well defined via the usual axioms for a state model). Now if $t_o \le t_1$ we have

$$Q(X(t_1)) = \bar{F}(a,t_1) = W(a,t_1) - \bar{D}(a,t_1) \le W(a,t_1) - \bar{D}(a,t_o) = W(a,t_1,t_o) + Q(X(t_o)) \qquad (6.7)$$

verifying that the network is dissipative and completing the proof.

Theorem: A network is dissipative in the sense of (6.1) if and only if $F(a,t) \ge 0$ for all a and t. The proof follows from Willems[7] and the above lemma.

The theorem clearly illustrates that the internal dissipation concept of Willems is weaker than our external concepts since a network with $D(a,t) < 0$ may be dissipative which is not the case if any of the passivity concepts studied in section IV hold. If we assume zero initial stored energy the Willems dissipation concept and the condition that the recoverable and unrecoverable energy are both positive coincide.

We say that a network is dissipative with zero initial energy if there exists an energy storage function Q mapping the state space to the non-negative real numbers such that equation (6.1) is satisfied and

$$\underset{t \to \infty}{\text{limit}} Q(X(t)) = 0 \qquad (6.9)$$

for all inputs a. Note that we do not require that the state itself go to zero (or for that matter that zero be in the state space) but only that the energy associated with the state go to zero at $-\infty$.

Theorem: A network is dissipative with zero initial energy if and only if

$$D(a,t) \ge 0 \quad , \; F(a,t) \ge 0$$

for all a and t.

Proof: If the network is dissipative with zero initial energy it is dissipative hence $F(a,t) \ge 0$ and

$$0 \le F(a,-\infty) = \underset{t \to -\infty}{\text{limit}}[F(a,t)] \; \underset{t \to -\infty}{\text{limit}} \; \underset{Q}{\text{minimum}}[Q(X(t))] \le \underset{t \to -\infty}{\text{limit}}[Q(t))] = 0 \qquad (6.10)$$

$$F(a,-\infty) = 0, \quad D(a,-\infty) = W(a,-\infty) = 0 \qquad (6.11)$$

and since $D(a,t)$ is non-decreasing in t we have

$$D(a,t) \geq D(a,-\infty) = 0 \tag{6.12}$$

as required. Conversely, if F and D are both non-negative we must have $F(a,-\infty) = 0$. Letting

$$Q(X(t)) = F(a,t) \tag{6.13}$$

where a is any input that drives the network to state $X(t)$ at time t the network is dissipative with zero initial energy.

CONCLUSIONS

Our purpose in the preceding has been to indicate the various alternative energy measures applicable to nonlinear networks and the passivity concepts which result therefrom. Unlike the linear case these passivity concepts are not equivalent and hence we treat them as equal concepts without distinguishing one as the definition of passivity. The close relationship between recoverable and unrecoverable energy and Willems' internal dissipation concept, however, tend to indicate that this formulation of the passivity concept may, in fact, be "more equal" than the others.

REFERENCES

1. Youla, D.C., L.J. Castriota and H.J. Carlin, Bounded Real Scattering Matrices and the Foundations of Linear Passive Network Theory, IRE Trans. on Circuit Theory, Vol. CT-6, pp 102-124, (1959).
2. Newcomb, R.W., LINEAR MULTIPORT SYNTHESIS, N.Y., McGraw-Hill, 1966.
3. Saeks, R., "Finite Energy Networks," IEEE Trans. on Circuit Theory, Vol. CT-17, pp. 618-619, (1970).
4. Winslow, L., and R.Saeks, "Lossless Nonlinear Networks," IEEE Trans. on Circuit Theory, Vol. CT-19, p. 392, (1972).
5. Gersho, A., "Nonlinear Systems with a Restricted Additivity Property," IEEE Trans., Vol. CT-16, pp. 150-154, (1969).
6. DeSantis, R., Ph.D. Dissertation, Univ. of Michigan, Ann Arbor,1971.
7. Willems, J.C., "Dissipative Dynamical Systems".
8. Saeks, R., GENERALIZED NETWORKS, New York, Holt, Rinehart and Winston, 1972.
9. DeCarlo, R.A., M.S. Thesis, Univ. of Notre Dame, Notre Dame, Ind. (in preparation).
10. Porter, W.A., "Some Circuit Theory Concepts Revisited," Int. J. on Control, Vol. 12, pp. 433-448, (1970).
11. Porter, W.A., "Nonlinear Systems in Hilbert Space", Int. J. on Control, Vol. 13, pp. 593-602, (1971).
12. Kalman, R.E., M.A. Arbib and P.L. Falb, TOPICS IN MATHEMATICAL SYSTEM THEORY, New York, McGraw-Hill, 1969.

SWITCHED ELECTRICAL NETWORKS AND BILINEAR EQUATIONS

DR. JONATHAN R. WOOD
Central Research Laboratories
Texas Instruments Incorporated
Dallas, Texas

The class of systems describable by a set of linear differential equations is restricted; most systems encountered in electrical power processing (i.e. DC-DC conversion, etc.) cannot be described by linear equations. In this paper we consider the question "What kinds of state equations arise in the description of power processing systems?" We outline the role played by Lie groups and Lie algebras in characterizing the inherent dynamical features of these systems, and give network examples.

1. BILINEAR EQUATIONS

Consider an electrical network composed of resistors, inductors, capacitors, transformers, batteries, current sources, and ideal switches. With these components we can model the essential features of power conversion networks. The state of this network will be a vector x(t) in R^n whose instantaneous value represents the inductor currents and capacitor voltages. If the switches are held fixed, the state equations take the linear form

$$\left\{ \dot{\underline{x}} = \underline{A}\,\underline{x} + \underline{B}\,\underline{u} \right. \qquad (1)$$

The eigenvalues of the constant matrix $\underline{A}$ are the natural frequencies of the circuit. If the switches are considered as controls, with position of switch i being given by u_i where $u_i = 0$ or 1, then $\underline{A}$ will be a function of the u_i's. The resulting equations take the form

$$\left\{ \dot{\underline{x}} = (\underline{A}_0 + u_1\underline{A}_1 + u_2\underline{A}_2 + \cdots + u_m\underline{A}_m)\underline{x} + (\underline{b}_1 u_1 + \underline{b}_2 u_2 + \cdots + \underline{b}_m u_m). \right. \qquad (2)$$

Now any system which can be realized by a set of equations of the form (2), i.e.

$$\left\{ \begin{aligned} \dot{\underline{z}} &= \Big(\underline{F}_0 + \sum_{i=1}^{m} u_i\underline{F}_i\Big)\,\underline{z} + \sum_{i=1}^{m} u_i\underline{g}_i \\ \underline{y} &= \underline{H}\,\underline{z} \end{aligned} \right. \qquad (3)$$

can be realized by a set of equations of the form

$$\left\{ \begin{aligned} \dot{\underline{x}} &= \Big(\underline{A}_0 + \sum_{i=1}^{m} u_i\underline{A}_i\Big)\,\underline{x} \\ \underline{y} &= \underline{C}\,\underline{x} \end{aligned} \right. \qquad (4)$$

To see this we let $\underline{A}_i$ be obtained by adding an extra row and column to $\underline{F}_i$, thus:

$$\underline{A}_0 = \begin{bmatrix} \underline{F}_0 & \underline{0} \\ \underline{0} & 0 \end{bmatrix} \qquad\qquad \underline{A}_i = \begin{bmatrix} \underline{F}_i & \underline{g}_i \\ \underline{0} & 0 \end{bmatrix},$$

and letting $\underline{x} = \begin{bmatrix} \underline{z} \\ 1 \end{bmatrix}$ and $\underline{C} = [\,\underline{H}\ \ \underline{0}\,]$.

We are therefore interested in equations in the form of (4). This is called the bilinear form. This form is even more general than we have just shown it to be, for it is shown in reference 2 that if $\underline{P}(\underline{x}) = \underline{P}(x_1, x_2, \cdots, x_n)$ is a multinomial expression in the variables $x_1, \cdots, x_n$, then any system of the form

$$\left\{ \begin{aligned} \dot{\underline{x}} &= \left(\underline{A}_0 + \sum_{i=1}^{m} u_i \underline{A}_i \right) \underline{x} \\ \underline{y} &= \underline{P}(\underline{x}) \end{aligned} \right. \tag{5}$$

can also be realized in the form of (4). Thus nonlinear output maps can be reduced to linear forms provided that they are of the finite power series type. Again, this is done by extending the dimension of the state vector. For example, given the equations

$$\left\{ \begin{aligned} \dot{x} &= -x + u \\ y &= x^3 \end{aligned} \right. ,$$

we would use a state vector $\underline{x}$ with elements $(1, x, x^2, x^3)$.

2. THE NATURE OF SOLUTIONS FOR BILINEAR EQUATIONS

For the linear equations (1) it is well known that the explicit solution for $\underline{x}(t)$ in terms of $\underline{u}(t)$ is given, for all t, by

$$\left\{ \underline{x}(t) = e^{\underline{A}t} \underline{x}(0) + \int_0^t e^{\underline{A}(t-\tau)} \underline{B}\, u(\tau) d\tau \right. . \tag{6}$$

This formula for $\underline{x}(t)$ is sometimes called the Variation of Constants Formula, and is useful not only for explicit calculation of solutions, but also for determining properties of these solutions.

It is not possible to write down an analogous explicit solution to equation (4) for all t. However, as shown in reference [1] we can study the intrinsic properties of the solutions of (4) by using the tools provided by the theory of matrix Lie Groups and Lie Algebras (Refs. [5], [6], [1], [7], [3]). To this end, we now introduce these concepts.

Let R^{nxn} denote the set of real nxn matrices; it is a vector space of dimension n^2. A Lie Algebra L in R^{nxn} is a vector subspace of R^{nxn} which has the property that if $\underline{X}$ and $\underline{Y}$ belong to L, then so does $[\underline{X}, \underline{Y}] = \underline{X}\underline{Y} - \underline{Y}\underline{X}$. We call $[\underline{X}, \underline{Y}]$ the commutator or Lie product of $\underline{X}$ and $\underline{Y}$. As examples of matrix Lie algebras we have:

Example 1: The set of all real nxn matrices R^{nxn} is itself a Lie algebra, sometimes called the general linear Lie algebra and denoted $g\ell(n, R)$.

Example 2: The set of all nxn real matrices with zero trace is called the special linear Lie algebra, and denoted sl(n, R).

Example 3: The set of all nxn real skewsymmetric matrices, i.e. those which satisfy $\underline{X}' + \underline{X} = \underline{0}$, is called the orthogonal Lie algebra and denoted o(n, R).

Example 4: The set of all 2n x 2n real symplectic matrices, i.e. those which satisfy $\underline{X}'\underline{J} + \underline{J}\,\underline{X} = 0$ where

$$\underline{J} = \begin{bmatrix} \underline{0} & -\underline{I} \\ \underline{I} & \underline{0} \end{bmatrix} ,$$

is called the symplectic Lie algebra and denoted sp(n, R).

Example 5: Consider the three matrices

$$\underline{R}_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \quad \underline{R}_y = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad \underline{R}_z = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

These are a basis for o(3, R), and we find that

$$[\underline{R}_y, \underline{R}_z] = \underline{R}_x, \quad [\underline{R}_z, \underline{R}_x] = \underline{R}_y, \quad [\underline{R}_x, \underline{R}_y] = \underline{R}_z.$$

Example 6: The affine algebra of the line, aff(1), consists of all real matrices of the form $\begin{bmatrix} u & v \\ 0 & 0 \end{bmatrix}$. The two matrices $\underline{X} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $\underline{Y} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ form a basis, and we find that $[\underline{X}, \underline{Y}] = \underline{Y}$.

Given an arbitrary subset of R^{nxn} we can add additional elements to it so as to imbed it in a Lie algebra. To obtain the smallest Lie algebra which contains a given set N we first add to N all linear combinations of elements of N so as to obtain a vector space N_1. Then we form all possible Lie products of elements in N_1 to obtain a set $[N_1, N_1]$. We add this to N_1 and obtain $N_2 = N_1 + [N_1, N_1]$. If N_2 is not contained in N_1 we $N_3 = N_2 + [N_2, N_2]$, etc. This process must stop in a finite number of steps since otherwise at each stage we increase the dimension of the vector space by at least one, and the dimension cannot exceed n^2. We call the Lie algebra thus obtained the Lie algebra generated by N, and we denote it $\{N\}_A$. For example, in Example 5 above the Lie algebra generated by any two of $\underline{R}_x, \underline{R}_y, \underline{R}_z$ is o(3, R).

Let L be a Lie algebra. The set [L,L] of all possible Lie products of elements in L is called the derived algebra of L and is denoted L'. The derived algebra of L' is denoted L". Continuing, we have the derived series:

$$L \supset L' \supset L'' \supset \quad \cdots \quad L^{(m)} \supset L^{(m+1)} \supset \quad \cdots \quad .$$

The Lie algebra L is said to be solvable if $L^{(m)}$ is zero for some m. We shall call L simple if L' = L. For example o(3) in Example 5 above is simple, while if L = aff(1) as in Example 6, then L' is spanned by Y and L" is zero, so that L is solvable. It is a fact that any Lie algebra can be decomposed into the semidirect sum of simple and solvable parts.

Having introduced matrix Lie algebras we next introduce the concept of a matrix Lie Group. If M is a set of nonsingular matrices in R^{nxn}, we let $\{M\}_G$ denote the multiplicative matrix group generated by M, i.e. it is the smallest set of matrices in R^{nxn} which contains M and which is closed under multiplication and inversion. If N is a linear subspace of R^{nxn}, then let P be the set of all matrices of the form $e^{\underline{X}_1} e^{\underline{X}_2} \cdots e^{\underline{X}_p}$ where $\underline{X}_i \in N$ for each i and p = 0, 1, 2, $\cdots$. P contains no singular matrices since for any matrix $\underline{x}_i$, $\det(e^{\underline{X}_i}) = e^{\operatorname{tr} \underline{X}_i} > 0$.

Since it is closed under multiplication and inversion, P is a group, and we write $P = \{\exp N\}_G$. It is an interesting and useful fact that $\{\exp N\}_G = \{\exp \{N\}_A\}_G$, (Ref. [1] Theorem 1).

If L is a Lie algebra then we call $\{\exp L\}_G$ the Lie group associated with L. For a full treatment of the relationship between Lie groups and Lie algebras the reader is referred to Refs. [7] and [3]. We now give the Lie groups which are associated with the six Lie algebras given above:

Example 1: The general linear group GL(n, R) consists of all real nxn invertible matrices.

Example 2: The special linear group SL(n, R) consists of all real nxn matrices with determinant 1.

Example 3: The orthogonal group 0(n, R) consists of the real nxn matrices which satisfy $X'X = I$. The special orthogonal group S0(n) consists of all matrices in both 0(n) and SL(n, R).

Example 4: The symplectic group Sp(n, R) consists of the real 2n x 2n matrices which satisfy $\underline{X}'\underline{J}\ \underline{X} = \underline{I}$ where

$$\underline{J} = \begin{bmatrix} \underline{0} & -\underline{I} \\ \underline{I} & \underline{0} \end{bmatrix}$$

Example 5: The group S0(3) consists of all real 3 x 3 matrices which have determinant +1 and satisfy $X'X = I$.

Example 6: The affine group of the line consists of all real 2 x 2 matrices whose second row is [0 1]. Any such matrix is of the form $\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}$, and represents a transformation $y = ax+b$ of the real line.

Now we can discuss the question of obtaining solutions to the equation

$$\left\{ \dot{\underline{x}} = \left(\underline{A}_0 + \sum_{i=1}^{m} u_i \underline{A}_i \right) \underline{x} \ , \ \underline{x}(0) = \underline{x}_0 \ \varepsilon \ R^n \ . \right. \tag{7}$$

It is well known that the solution to (7) is given by

$$\left\{ \dot{\underline{x}}(t) = \underline{\Phi}(t)\, \underline{x}_0 \right. \tag{8}$$

where $\underline{\Phi}(t)$ is an nxn matrix, called the transition matrix for (7), which is the solution of

$$\left\{ \dot{\underline{\Phi}}(t) = \left(\underline{A}_0 + \sum_{i=1}^{m} u_i \underline{A}_i \right) \underline{\Phi} \ , \ \underline{\Phi}(0) = \underline{I} \ . \right. \tag{9}$$

We are, therefore, interested in solutions of equation (9). Now it can be shown (Ref. [1], Theorem 5) that for all t, $\underline{\Phi}(t)$ belongs to the matrix Lie Group $\{\exp\{\underline{A}_i\}_A\}_G$, that is, the matrix $\underline{\Phi}(t)$ will evolve on the Lie Group associated with the Lie Algebra generated by the coefficient matrices $\underline{A}_0, \underline{A}_1, \cdots, \underline{A}_m$.

There are two important consequences of this fact for power conversion networks described by bilinear equations. The first of these stems from the fact that in addition to having the properties of a group, a Lie Group has the properties of a manifold, that is, a subset of Euclidean space R^d with special geometrical characteristics. The two-dimensional surface of a sphere in three-dimensional space is an example of a manifold. The geometrical characteristics of the particular manifold on which the state of a bilinear system evolves will play a fundamental role in determining the nature of the behavior of the system. For example, if the Lie group is compact, i.e. closed and bounded as a subset of R^{nxn}, then we know that the state of the system is bounded, i.e. the amount of energy stored in the inductors and capacitors will be finite. For instance, the Lie group S0(n) is bounded, and the group Sp(n) is not. Thus, given a bilinear system of equations, the natural first question we ask is "On what Lie group does its state transition matrix evolve?"

The second consequence is that questions about the system (such as controllability, reachability, observability, and stabilizability) can be reduced to questions about the Lie algebra generated by the coefficient matrices. In many cases it is possible to arrive at conditions which are easily visualized and tested, as for example the results of reference [1]. Of relevance here is the classification result from the theory of Lie algebras, by which any simple Lie algebra is shown to be equivalent to one of a short list of canonical algebras, (Ref. [6], Chapter 2). For solvable algebras there is no such complete list of all possibilities, although a good many facts are known. Note that if $\underline{P}$ is a nonsingular matrix, equation (7) is unchanged by the change of variable $\underline{x} \to \underline{x}\,\underline{P}$, and equation (9) is unchanged by $\underline{\Phi} \to \underline{\Phi}\,\underline{P}$.

Now let us further discuss the problem of obtaining a solution to equation (9). Let $\{\underline{X}_1, \underline{X}_2, \cdots \underline{X}_k\}$ be a basis for the Lie algebra L generated by $\{\underline{A}_0, \underline{A}_1, \cdots, \underline{A}_m\}$. In reference [4] Magnus showed that there exists a t_0 such that for $0 \le t < t_0$ the solution to (9) can be expressed in the form

$$\{\ \underline{\Phi}(t) = e^{b_1(t)\underline{X}_1 + b_2(t)\underline{X}_2 + \cdots + b_k(t)\underline{X}_k} \tag{10}$$

where $b_1, \cdots, b_k$ are scalar functions of time satisfying differential equations which depend on L and $u_1, \cdots, u_m$. The difficulties with this approach to obtaining a solution are that it is difficult to derive and solve the differential equations for $b_1, \cdots, b_k$, and that only in severely restricted cases is the representation (10) valid for all t in $[0, \infty)$.

Wei and Norman (Ref. [8]) showed that it is often preferable to look for a solution in the form of a product of exponentials. In fact they showed that there exists a t_0 such that for $0 \le t < t_0$ the solution to (9) can be expressed in the form

$$\{\ \underline{\Phi}(t) = e^{g_1(t)\underline{X}_1}\, e^{g_2(t)\underline{X}_2} \cdots e^{g_k(t)\underline{X}_k} \tag{11}$$

where $g_1, g_2, \cdots, g_k$ are scalar functions of time satisfying differential equations which depend on L and on $u_1, \cdots u_m$. Moreover, if L is solvable, or if the matrices are 2×2, then this representation is global, i.e. it holds for all $0 \le t < \infty$.

In conclusion we note that while in some cases it may be possible to analyze the behavior of electrical power conversion networks with piecewise-linear models, concepts from the theory of Lie groups and Lie algebras are useful in characterizing the inherent dynamical features of such systems, especially since the methods and conclusions are basis-free, i.e. they do not depend on the particular basis chosen for the state space.

3. NETWORK EXAMPLES

(a) An SO(3) Network

Fig. 1 shows a simple network in which charge stored on one capacitor can be transferred by means of the inductor to the other capacitor. If the capacitances are different a voltage conversion will be effected. A transferral cycle might be executed as follows. Starting from $V_1(0) = V_{10}$, $I_3(0) = 0$, $V_2(0) = 0$, the switch is held in the $u = 0$ position until $V_1 = 0$ and all the energy is stored in L_3, at which time the switch is changed to the $u = 1$ position. It is held there until I_3 again becomes zero, at which time the switch is reverted to the $u = 0$ position. All the initial energy is now stored in C_2, with $V_2 = -\frac{C_1}{C_2} V_{10}$. While this network is too simple to model a complete DC-DC converter, it may very well model the conversion portion of a convertor in which the charge on C_1 is replenished from an external supply during the $u = 1$ portion of the cycle, and a load current is drawn from C_2 during the $u = 0$ portion. The equations for this network are

$$C_1 \dot{V}_1 = (1-u) I_3$$
$$C_2 \dot{V}_2 = u I_3$$
$$L_3 \dot{I}_3 = -(1-u) V_1 - u V_2 \ .$$

Letting $V_i = \frac{x_i}{\sqrt{C_i}}$ for $i = 1, 2$ and $I_3 = \frac{x_3}{\sqrt{L_3}}$ so that $\frac{1}{2}(\underline{x}'\underline{x})$ is the total stored energy, we obtain

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \left\{ \begin{bmatrix} 0 & 0 & \omega_1 \\ 0 & 0 & 0 \\ -\omega_1 & 0 & 0 \end{bmatrix} + u \begin{bmatrix} 0 & 0 & -\omega_1 \\ 0 & 0 & \omega_2 \\ \omega_1 & -\omega_2 & 0 \end{bmatrix} \right\} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

where $\omega_1 = \frac{1}{\sqrt{L_3 C_1}}$, $\omega_2 = \frac{1}{\sqrt{L_3 C_2}}$,

i.e. $\dot{\underline{x}} = \left[(1-u) \underline{A}_1 + u \underline{A}_2 \right] \underline{x}$

where $\underline{A}_1 = \begin{bmatrix} 0 & 0 & \omega_1 \\ 0 & 0 & 0 \\ -\omega_1 & 0 & 0 \end{bmatrix}$ $\quad \underline{A}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \omega_2 \\ 0 & -\omega_2 & 0 \end{bmatrix}$.

We see that $\underline{A}_1$ and $\underline{A}_2$ are just scalar multiples of $\underline{R}_y$ and $\underline{R}_x$ of the Lie algebra Example (5). Thus the Lie algebra generated by $\underline{A}_1$ and $\underline{A}_2$ is o (3), and the transition matrix $\underline{\Phi}(t)$ for this network evolves on the Lie group SO(3). The state vector $\underline{x}(t)$ evolves on the 2-sphere S^2.

(b) Simple and Solvable Parts

If we add a current sink in parallel with C_2 of Fig. 1, to represent a load for instance, we obtain the network of Fig. 2.

The state equations are now

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \left\{ \begin{bmatrix} 0 & 0 & \omega_1 \\ 0 & 0 & 0 \\ -\omega_1 & 0 & 0 \end{bmatrix} + u \begin{bmatrix} 0 & 0 & -\omega_1 \\ 0 & 0 & \omega_2 \\ \omega_1 & -\omega_2 & 0 \end{bmatrix} \right\} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ \gamma \\ 0 \end{bmatrix}$$

where $\gamma = \frac{I}{\sqrt{C_2}}$.

Letting $x_4 = 1$ we can put this in bilinear form, thus:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \left\{ \begin{bmatrix} 0 & 0 & \omega_1 & 0 \\ 0 & 0 & 0 & \gamma \\ 0 & -\omega_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + u \begin{bmatrix} 0 & 0 & -\omega_1 & 0 \\ 0 & 0 & \omega_2 & 0 \\ \omega_1 & -\omega_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right\} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

i.e. $\quad \dot{\underline{x}} = (\underline{B}_0 + u\underline{B}_1)\underline{x}$.

The Lie algebra L generated by $\underline{B}_0$ and $\underline{B}_1$ is six-dimensional and has as a basis the following matrices:

$$\underline{X}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{X}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{X}_3 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\underline{X}_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{X}_5 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{X}_6 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} .$$

We see that $\underline{X}_1, \underline{X}_2, \underline{X}_3$ are a reducible representation of o(3), while $\underline{X}_4, \underline{X}_5, \underline{X}_6$ represent a solvable Lie algebra. Thus L is decomposable into simple and solvable parts, with the simple part determined by the "natural dynamics" of the network, i.e. the interconnection of its inductors and capacitors, while the solvable part is contributed by the "driving forces", i.e. the batteries and current sources. This type of decomposition is a general characteristic of the types of networks we are considering.

(c) A Reducible Network

Our final example shows how reducibility of the Lie algebra can correspond to reducibility of the network. In Fig. 3 we wish to transfer energy from the battery E to the output capacitor C_2. We might ask whether an unlimited amount of energy can be extracted from the battery. The two switches are operated synchronously. In a similar manner as before we obtain the state evolution equations as

$$\dot{\underline{x}} = (\underline{A}_0 + u\underline{A}_1)\,\underline{x}$$

where

$$\underline{A}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \omega_2 & 0 \\ 0 & -\omega_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \underline{A}_1 = \begin{bmatrix} 0 & 0 & -\omega_1 & 0 \\ 0 & 0 & -\omega_2 & 0 \\ \omega_1 & \omega_2 & 0 & -\gamma \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $\quad \omega_1 = \dfrac{1}{\sqrt{L_3C_1}}, \quad \omega_2 = \dfrac{1}{\sqrt{L_3C_2}}, \quad \gamma = \dfrac{E}{\sqrt{L_3}}$.

The Lie algebra generated by A_0 and A_1 is o (3), and thus we conclude that this network is dynamically similar to that of Fig. 1 and that only a finite amount of energy can be stored in C_1, C_2, L_3 since the associated Lie group is bounded. This may at first seem surprising, since the network of Fig. 3 has a battery in it. In the case $\omega_1 = \omega_2 = 1$ the Lie algebra has a basis

$$\underline{X}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{X}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & \gamma \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \underline{X}_3 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -\gamma \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $[\underline{X}_2, \underline{X}_3] = \underline{X}_1, \; [\underline{X}_3, \underline{X}_1] = \underline{X}_2, \; [\underline{X}_1, \underline{X}_2] = \underline{X}_3$.

Examination of the network of Fig. 3 shows that we can reduce it to a simpler form. The network is unchanged if the switch u and capacitor C_1 are interchanged, giving the network of Fig. 1 but for a battery in series with C_1. Now the total energy which can be extracted from the combination of Fig. 4 is $\frac{1}{2} CE^2$. (This is an interesting singular optimal control problem which may be solved by observing that the arrangement of Fig. 4 is externally identical with that of Fig. 5, since both are governed by

$$V(t) = E + \frac{1}{C} \int_0^t I(\tau)\, d\tau \quad . \quad)$$

In fact, we see that the network of Fig. 3 is essentially the network of Fig. 1, but with the capacitor C_1 being given an extra initial voltage of E. The matrices $\underline{X}_1$, $\underline{X}_2$, $\underline{X}_3$ given here are a reducible representation of the Lie algebra o (3): they can be reduced by eliminating from each one the last row and column, which represents the contribution of the batteries and current sources in the network, as we saw in Example (b).

ACKNOWLEDGEMENTS

This paper is based on a doctoral dissertation at the Massachusetts Institute of Technology, which was supported in part by NASA under Grant NGR-22-007-172. The author is grateful to Professor Roger W. Brockett of Harvard University for help and guidance in this work.

REFERENCES

[1] R.W. Brockett, "System Theory on Group Manifolds and Coset Spaces", SIAM J. on Control, Vol. 10, No. 2, May 1972, pp. 265-284.
[2] R.W. Brockett, "On the Algebraic Structure of Bilinear Systems", in Theory and Applications of Variable Structure Systems, R. Mohler and A. Ruberti Eds., Academic Press, New York, 1972.
[3] C. Chevalley, Theory of Lie Groups, Princeton University Press, Princeton, 1946.
[4] W. Magnus, "On the Exponential Solution of Differential Equations for a Linear Operator", Comm. Pure Appl. Math., Vol. 7, 1954, pp. 649-673.
[5] L. Markus, "Dynamical Systems on Group Manifolds", in Differential Equations and Dynamical Systems, J.K. Hale and J.P. LaSalle Eds., Academic Press, New York, 1967, pp. 487-493.
[6] H. Samelson, Notes on Lie Algebras, Van Nostrand, New York, 1969.
[7] F.W. Warner, Foundations of Differentiable Manifolds and Lie Groups, Scott Foresman, Glenview, Illinois, 1971.
[8] J. Wei and E. Norman, "On Global Representations of the Solutions of Linear Differential Equations as a Product of Exponentials", Proc. Am. Math. Soc., April 1964, pp. 327-334.
[9] J.R. Wood, Power Conversion in Electrical Networks, Ph.D. Thesis, M.I.T., Cambridge, Massachusetts, 1973.

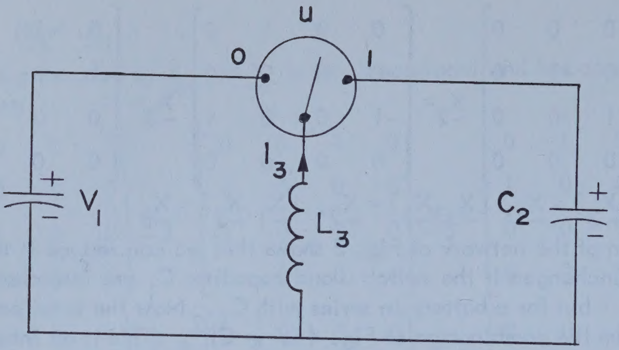

u
0
1
I3
C1
+
−
V1
L3
C2
+
−
V2

Fig 1

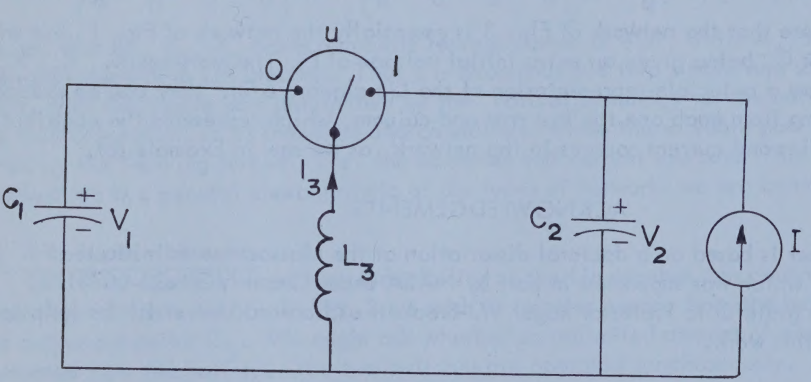

u
O
1
I3
C1
+
−
V1
L3
C2
+
−
V2
I

Fig 2

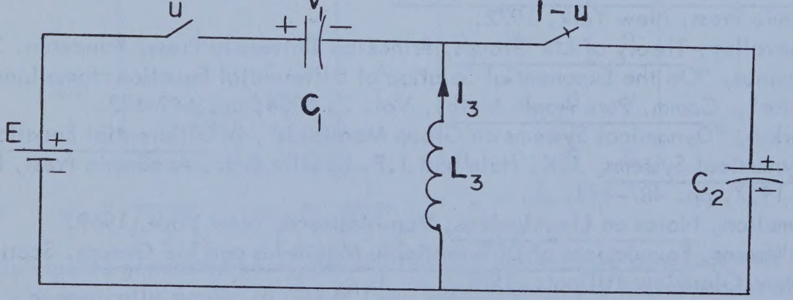

u
V1
+
−
1−u
C1
I3
E
+
−
L3
C2
+
−
V2

Fig 3

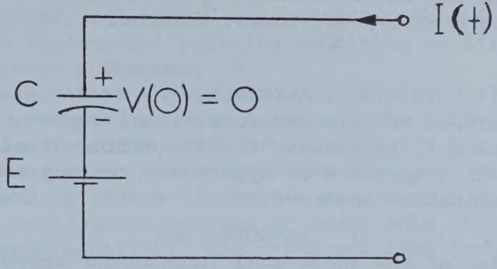


Fig 4

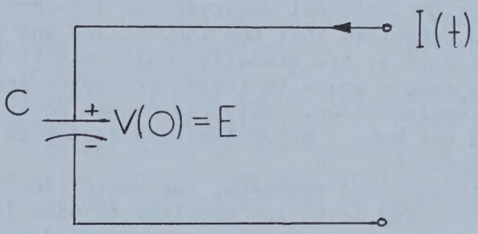


Fig 5

A GENERALIZED REPRESENTATION THEOREM WITH APPLICATIONS TO ESTIMATION AND HYPOTHESIS TESTING

KEH-PING DUNN AND IAN B. RHODES
Washington University, St. Louis, Missouri

ABSTRACT

The simple notion of relative conditional expectation is combined with Girsanov's Theorem and, in places, the Ito differential rule to provide a direct and attractive approach to hypothesis testing and detection, estimation, and representation theorems.

I. INTRODUCTION

The basic idea of the successful innovations approach to estimation and detection problems is to replace the observation process by a new process constructed from it. This new process, the innovations process, is a Wiener process, and because of the special characteristics of Wiener processes the resulting estimation and detection problems are greatly simplified. The innovations process is constructed by changing the functional relationship between the underlying sample space and the space of trajectories of the random processes, but leaving the probability measure on the underlying sample space unchanged. In order to ensure that substituting the innovations for the original observations does not change the problem, it is necessary to establish that the innovations and observations carry the same information, i.e. are causally equivalent in the sense that the two functions from sample space to trajectory space are related by a causally-invertible transformation. This has proved to be a difficult problem in some cases, and has been a major limiting factor in the applicability of the innovations approach.

As with the innovations approach, the basic idea in this paper is also to convert estimation and detection problems to ones involving a Wiener process (or a standard Poisson process) constructed from the observations. In contrast to the innovations approach, however, this process is constructed by leaving unchanged the functional relation between the underlying sample space and the observation trajectories, and changing instead the measure on the sample space using standard measure transformation techniques provided by Girsanov's Theorem [1]. Because we are considering the same functional relation under two different, but equivalent, probability measures, questions of causal equivalence never arise. The limiting factor on applicability here is the equivalence of two measures, a problem for which a large literature already exists. What does remain is to relate the conditional expectations under the new and old measures. This, however, is immediately taken care of by the simple and standard notation of *relative conditional expectation* [2].

In the remainder of this paper we illustrate the simplicity and versatility of these ideas by considering their application to a sampling of prototype estimation and detection problems. In each case we obtain the most general known results, or extensions of these, by particularly simple and straightforward calculations that are usually shorter and more direct than existing derivations. At the same time, this approach provides insights and interpretations not otherwise available.

Throughout the paper we treat simultaneously two observation models:

This work was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, under Grant AFORS 73-2496, and in part by the National Science Foundation under Grants GK-22905A, GI-34969X.

(i) the familiar "signal in additive white Gaussian noise" model

(ii) observed is a counting process whose intensity is signal-dependent. In addition to providing a unified treatment of these two observation types, our approach also gives a unifying viewpoint on representation theorems, state estimation problems, and hypothesis-testing problems, since all emerge as special cases of a fundamental relation combining relative conditional expectation with Girsanov's Theorem.

Insofar as representation theorems for the conditional distribution of the signal given the observations are concerned, the notion of relative conditional expectation in conjunction with Girsanov's Theorem provides what might be thought of as a version of Bayes Rule that is especially well-suited to problems involving stochastic processes. This is an area in which there has been no rigorous version of Bayes Rule of satisfactorily wide generality and applicability. The representation theorems we derive unify and generalize all existing results of this type, and obstacles to further generalization are clearly delineated.

We merely sketch the application of our approach to the derivation of stochastic integral and differential equations for the conditional mean of a function of the state by considering just two examples (fixed-point smoothing for the counting-process observation model, and correlated-noise filtering). Our principal purpose here is to indicate the versatility of this approach in providing rigorous but nevertheless short and simple derivations of results of this type. In particular, it provides an attractive alternative to the innovations approach, both in viewpoint and the simplicity of calculations. Our consideration of hypothesis-testing and detection problems is in the same spirit. We derive formulas for the posterior probabilities and likelihood ratios connected with the M-ary hypothesis testing problem. The specialization of these to M=2 gives a particularly simple and direct proof of the "generalized likelihood formula" [3], [4] for the detection problem, and provides a different viewpoint and considerable insight into this class of problems.

II. PRELIMINARIES

Let Θ be a random variable on the probability space $(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$, $\tilde{P} << P$ a probability measure on $(\Omega, \mathcal{F})$, and L the Radon-Nikodym derivative of $\tilde{P}$ with respect to P. Then the *relative conditional expectation* $\tilde{E}[\Theta|\mathcal{G}]$ of Θ given $\mathcal{G}$ with respect to $\tilde{P}$ is a $\mathcal{G}$-measurable random variable defined up to $\tilde{P}_{\mathcal{G}}$-equivalence by [2]

$$\int_G E[\Theta|\mathcal{G}]\, d\tilde{P}_{\mathcal{G}} = \int_G \Theta\, d\tilde{P} \quad \forall G \in \mathcal{G}$$

or, equivalently,

$$\tilde{E}[\Theta|\mathcal{G}] = E[\Theta L|\mathcal{G}] \Big/ E[L|\mathcal{G}] \quad \tilde{P}_{\mathcal{G}} \text{ - a.s.} \qquad \text{(RCE1)}$$

where $\tilde{P}_{\mathcal{G}}$ denotes the restriction of $\tilde{P}$ to $\mathcal{G}$.

In the application of (RCE1) to estimation and detection problems, we will be concerned with stochastic processes defined on a fixed measurable space $(\Omega, \mathcal{F})$. All processes are considered to be $\mathcal{F}_t$-adapted, where $\{\mathcal{F}_t, t \in [0,T]\}$ is an increasing family of sub-σ-algebras of $\mathcal{F}$, and defined on the fixed interval [0,T]. We denote by $z \overset{\Delta}{=} \{z_t, \mathcal{F}_t, P;\, t \in [0,T]\}$, or simply $\{z_t, \mathcal{F}_t, P\}$, an $\mathcal{F}_t$-adapted, P-integrable stochastic process on $(\Omega, \mathcal{F})$ under the probability measure P and taking values in $(S_z, \mathcal{B}_z)$. The same mapping $z.(\cdot)$ from Ω into the function space of trajectories on [0,T] but under the probability measure P^* will be denoted z^* or $\{z_t, \mathcal{F}_t, P^*\}$. We denote by $\sigma(z_t)$ the sub-σ-algebra of $\mathcal{F}_t$ generated by the random variable z_t, and by $\mathcal{Z}_t$ the sub-σ-algebra of $\mathcal{F}_t$ generated by the collection of random variables $\{z_s; s \in [0,t)\}$, i.e.

$\sigma(z_t) = \bigvee_{B\in\mathcal{B}_Z}\{\omega : z_t(\omega)\in B\}$, $Z_t = \bigvee_{s<t}\sigma(z_s)$ where $\bigvee$ means "the least σ-algebra generated by".

Throughout the paper, $X = \{x_t, F_t, P\}$ will denote a "signal" process and $Z = \{z_t, F_t, P\}$ will be an "observation" process. Attention will be restricted to two types of observation processes:

Additive Wiener process model: Z satisfies the Ito equation

$$dz_t = g(x_t,t)dt + G(t)dw_t \; ; \; z_0(\omega) = 0 \tag{W}$$

where $W = \{w_t, F_t, P\}$ is a Wiener process taking values in R^m, $G(t): R^m \to R^m$ is invertible, g is jointly measurable in x_t and t, and

$$\int_0^T ||G(t)||^2 dt < \infty , \int_0^T ||g(x_t,t)||^2_{W^{-1}(t)} \; dt < \infty \quad \text{P-a.s.}$$

where $W(t) = G(t)G'(t)$ and $||\;\;||$ denotes any appropriate norm.

Counting process model: Here Z satisfies

$$dz_t = dN_t \quad ; \quad z_0(\omega) = 0 \tag{C1}$$

where $N = \{N_t, F_t, P\}$ is a counting process with intensity $\Lambda = \{\lambda_t, F_t, P\}$ [3] satisfying

$$\lambda_t = g(x_t,t) \tag{C2}$$

where g is jointly measurable in x_t and t, $\int_0^T g(x_s,s)ds < \infty$ P-a.s., and g is such that Λ is predictable and positive a.s.

Remark: Our subsequent results easily extend to the case where the function g in (W) and (C2) depends on the entire past of X, i.e. for $g(\Pi_t x,t)$ where $(\Pi_t x)_s = x_s$ if $s\in[0,t]$ and $(\Pi_t x)_s = x_t$ if $s\in[t,T]$.

For some of the subsequent results we assume, in addition, that Z is doubly-stochastic with respect to X. For the observations (W), Z is *doubly stochastic* with respect to X if

$\{E[z_t - \int_0^t g(x_s,s)ds|X_t]$, F_t, P} is a Wiener process with covariance $\int_0^t W(s)ds$. Clearly, Z is doubly stochastic with respect to X if X is independent of W. For the observations (C), Z is *doubly-stochastic* with respect to X if $\{E[z_t|X_t], F_t, P\}$ is a Poisson process with deterministic intensity $E[g(x_t,t)|X_t]$ [3], [5]. This definition excludes self-exciting processes where X(or, more generally, Λ) depend on Z.

Combined with the definition of relative conditional expectation, Girsanov's Theorem [1] plays a central role in the ensuing developments. The following versions, one for each observation model, are specialized with a view to subsequent purposes and are stated without proof.

Theorem GW (Girsanov) Let $Z = \{z_t, F_t, P\}$ satisfy (W). Suppose one of the following is true

i) Z is doubly stochastic with respect to X;

ii) for some $K < \infty$ $\int_0^T ||g(x_t,t)||^2_{W^{-1}(t)} \; dt < K$ P-a.s.;

iii) for some $\delta > 0$, $E\left[\exp\left\{(1+\delta)\int_0^T ||g(x_t,t)||^2_{W^{-1}(t)}\, dt\right\}\right] < \infty$.

Now let P and L_t be defined by

$$dP^* = \exp\left\{-\int_0^T g_t' G^{-1}(t)\,dw_t - \frac{1}{2}\int_0^T ||g_t||^2_{W^{-1}(t)}\,dt\right\} \quad dP \tag{GW1}$$

$$L_t = \exp\left\{\int_0^t g_s' \, W^{-1}(s)\,dz_s - \frac{1}{2}\int_0^t ||g_s||^2_{W^{-1}(s)}\,ds\right\} \tag{GW2}$$

where $g_t = g(x_t,t)$. Then

a) $P(\Omega) = 1$, i.e., P is a probability measure

b) $Z^* = \{z_t, \mathcal{F}_t, P^*\}$ is a Wiener process with covariance $\int_0^t W(s)\,ds$

c) $P^* \equiv P$, i.e., $P^* \ll P$ and $P \ll P^*$

d) $dP = L_T dP^*$

e) $\{L_t, \mathcal{F}_t, P^*\}$ is a martingale; $E^*[L_t | \mathcal{F}_s] = L_s$, $s \leq t$, P^*-a.s.

Theorem GC (Girsanov) Let $Z = \{z_t, \mathcal{F}_t, P\}$ satisfy (C1). Suppose $g(x_t,t)$ is uniformly bounded on t P-almost surely and let P and L_t be defined by

$$dP^* = \exp\left\{-\int_0^T \ln g_t \, dz_t + \int_0^T (g_t - 1)\,dt\right\} \quad dP \tag{GC1}$$

$$L_t = \exp\left\{\int_0^T \ln g_s \, dz_s - \int_0^t (g_s - 1)\,ds\right\} \tag{GC2}$$

Then a) $P^*(\Omega) = 1$, i.e., P^* is a probability measure

b) $Z^* = \{z_t, \mathcal{F}_t, P^*\}$ is a standardized Poisson process ($\lambda_t \equiv 1$)

c) $P^* \equiv P$, i.e., $P^* \ll P$ and $P \ll P^*$.

d) $dP = L_T \, dP^*$

e) $\{L_t, \mathcal{F}_t, P\}$ is a martingale, i.e., $E^*[L_t | \mathcal{F}_s] = L_s$, $s \leq t$, P^*-a.s.

Corollary G: If, in addition to the hypotheses of Theorem GW or GC, Z is a doubly stochastic process with respect to X, then

a) Under P^*, X^* is completely independent of Z^*

b) $P^*_X = P_X$, i.e. the restrictions of P^* and P to X coincide.

Assuming the hypotheses of either of the two Girsanov Theorems are met, let Z_t play the role of G, P of $\tilde{P}$, and P^* of P in the definition (RCE1) of relative conditional expectation. Because $P^* \equiv P$ and $dP = L_T \, dP^*$, (RCE1) becomes

$$E[\Theta | Z_t] = E^*[\Theta L_T | Z_t] \Big/ E^*[L_T | Z_t] \qquad P_{Z_t}\text{-a.s.} \tag{RCE2}$$

In our subsequent applications, Θ takes one of three specific forms:

Detection and Hypothesis Testing: $\Theta = I_H$, $H \in \mathcal{F}_0$

Representation Theorems: $\Theta = I_A$, $A = \{\omega: x_s(\omega) \in B, \ B \in \mathcal{B}_x\}$

Estimation: $\Theta = h(x_s)$, where h has all required technical properties

where I_A is the indicator function of the set A. In the first case, Θ is trivially $\mathcal{F}_t$-measurable. In the last two cases Θ is $\sigma(x_s)$-measurable, and therefore $\mathcal{F}_s$-measurable, and, if $s \leq t$ (filtering and smoothing), Θ is trivially also $\mathcal{F}_t$-measurable. Under these circumstances it is easy to show that L_T can be replaced by L_t in (RCE2). The same simplification can also be made for $s > t$ (prediction) if Z is doubly-stochastic with respect to X; the proof is omitted.

Lemma G: Under the assumptions of either Theorem GW or GC, (RCE2) can be replaced by

$$E[\Theta | Z_t] = E^*[\Theta L_t | Z_t] \Big/ E^*[L_t | Z_t] \qquad \text{a.s.} \tag{RCE3}$$

if either of the following conditions is satisfied:

(i) Θ is $\mathcal{F}_t$-measurable

(ii) Θ is $\sigma(x_s)$-[or $\mathcal{X}_s$-] measurable and Z is doubly-stochastic with respect to X.

Equation (RCE3) is the basis of our subsequent applications. Notice that it gives a relation between conditional expectations under P and P*, and conditional expectations under P* given $\mathcal{Z}^*$ $(=\mathcal{Z})$ are much more easily calculated because, recalling Theorems GW and GC, Z* is a Wiener process or standard Poisson process under P*. Armed with just (RCE3) and the Ito differential rule, we now examine a number of hypothesis-testing, detection, and estimation problems, and indicate the simplicity and directness with which known results can be rederived and generalizations of them made.

III. DETECTION AND HYPOTHESIS-TESTING

We first study M-ary detection problems involving signals observed from the observation models (W) or (C). The signal process X in this section is considered to have the form $x_t = \sum_{i=0}^{M-1} I_{H_i} x_t^i$, where $\{H_i\}$ are M mutually disjoint $\mathcal{F}_0$-measurable sets in Ω with $\bigcup_{i=0}^{M-1} H_i = \Omega$, and $X^i = \{x_t^i, \mathcal{F}_t, P\}$, i = 0,1, ..M-1, are M distinct $\mathcal{F}_t$-measurable signal processes such that Theorem GW or GC holds for X, which precludes singular problems. We do not require that Z is doubly-stochastic with respect to X. Then whenever $\omega \in H_i$ we have $x_t = x_t^i$, and this is equivalent to the statement "the hypothesis $(h_i : x_t = x_t^i)$ is true". Now let the a priori probability that h_i is true be $P(H_i | \mathcal{Z}_0) \triangleq P_i$; of course, and $\sum_{i=0}^{M-1} P_i = 1$. To avoid trivialities, assume $P_i > 0$ $\forall i = 0,1,..,M-1$. The "normalized" likelihood ratio Λ_t^{ij}, of hypothesis h_i versus h_j at time $t \in [0,T]$ is defined by

$$\Lambda_t^{ij} \triangleq P(H_i | \mathcal{Z}_t) \Big/ P(H_j | \mathcal{Z}_t) \tag{D1}$$

Remark: The relation between Λ_t^{ij} and the "traditional" likelihood ratio $\bar{\Lambda}_t^{ij}$ [6] is

$$\bar{\Lambda}_t^{ij} = (P_j | P_i)\, \Lambda_t^{ij} \qquad t \in [0,T] \tag{D2}$$

One reason we prefer (D1) is because it explicitly gives the ratio of the conditional probabilities that hypothesis h_i is true versus hypothesis h_j is true given the observation up to time t. Another advantage is the ease with which we can obtain the well-known general likelihood formula [4], [6]-[8] for $\bar{\Lambda}_t^{ij}$ via simple Ito calculus.

To calculate $\bar{\Lambda}_t^{ij}$, we will calculate the differential equation for $P(H_i | \mathcal{Z}_t) \triangleq P_t^i$ first, then use the Ito differential again on (D1). Since I_{H_i} is $\mathcal{F}_t$-measurable, we can let $\Theta = I_{H_i}$ in (RCE3) and notice that $g_t \triangleq g(x_t,t) = g\left(\sum_{i=0}^{M-1} I_{H_i} x_t^i, t\right) = \sum_{i=0}^{M-1} I_{H_i} g(x_t^i, t)$; simple algebra then yields:

Observations (W):

$$dP_t^i = P_t^i(\hat{g}_{it} - \bar{g}_t)' W^{-1}(t)(dz_t - \bar{g}_t dt); \quad P_0^i = P_i \tag{D3}$$

$$d\ln \Lambda_t^{ij} = \tfrac{1}{2}[\hat{g}_{it} - \hat{g}_{jt}]' W^{-1}(t)(d\nu_t^i + d\nu_t^j); \quad \Lambda_0^{ij} = P_i/P_j \tag{D4}$$

where $g_i \triangleq g(x_t^i, t)$, $\bar{g}_t \triangleq \sum_{i=0}^{M-1} P_t^i \hat{g}_{it}$

$\hat{g}_{it} \triangleq E[g_t | Z_t \cap H_i] = E^*[g_t I_{H_i} L_t | Z_t] \big/ E^*[I_{H_i} L_t | Z_t]$,

$d\nu_t^i \triangleq dz_t - \hat{g}_{it} dt$ and $Z_t \cap H_i \triangleq \{A \cap H_i ;\ A \in Z_t\}$

Remark: Let $M=2$ and $g_0=0$, $g_1=g(x_t,t)$ in (D4). We have

$$d\ln \Lambda_t^{10} = \tfrac{1}{2}[\hat{g}_t - 0]' \ W^{-1}(t)[2dz_t - \hat{g}_t dt] \ ; \quad \Lambda_0^{10} = P_1/P_0$$

Integrating this and with simple algebra yields

$$\bar{\Lambda}_t^{10} = \exp \left\{ \int_0^t \hat{g}_s' \ W^{-1}(s) dz_s - \frac{1}{2} \int_0^t \ ||\hat{g}_s||^2_{W^{-1}(s)} \, ds \right\} \tag{D5}$$

This expression (D5) has been obtained by several authors, in particular, by [6] under a restricted class of signal processes, and by [4] under effectively the same assumptions as those we have used. The expression (D4) is given in [9] under the assumptions of [4]. The derivation of (D5) in [4] is a lengthy one that involves the (conditional) innovations process and appeals to results from martingale theory, both of which are avoided in our approach. Like us, [4] makes use of a version of Girsanov's theorem, but our different viewpoint leads to a short, direct derivation that needs in addition only the definition of relative conditional expectation and simple, straightforward calculations using the Ito differential rule. Furthermore, this approach provides an equally simple derivation of the counterpart of this result for the counting process observation model.

Observations (C): Here the analogous formulas for (D3) and (D4) are

$$dP_t^i = P_t^i \ (\hat{g}_{it} - \bar{g}_t) \bar{g}_t^{-1} (dz_t - \bar{g}_t dt) \ ; \ P_0^i = P_i \tag{D6}$$

$$d\ln \Lambda_t^{ij} = \ln(\hat{g}_{it} \big/ \hat{g}_{jt}) dz_t - [\hat{g}_{it} - \hat{g}_{jt}] dt \ ; \ \Lambda_0^{ij} = P_i | P_j \tag{D7}$$

Remark: Letting $M=2$ and $g_0 = \lambda_0(x_t,t)$, $g_1 = \lambda_1(x_t,t)$ in (D7) gives

$$d\ln \Lambda_t^{10} = \ln(\hat{\lambda}_{1t} | \hat{\lambda}_{0t}) dz_t - [\hat{\lambda}_{1t} - \hat{\lambda}_{0t}] dt \ ; \ \Lambda_0^{10} = P_1 | P_0$$

Integrating this and simple algebra yields

$$\bar{\lambda}_t^{10} = \exp \left\{ \int_0^t \ln(\hat{\lambda}_{1s} \big/ \hat{\lambda}_{0s}) dz_s - \int_0^t (\hat{\lambda}_{1s} - \hat{\lambda}_{0s}) ds \right\} \tag{D8}$$

The expression (D8) has been given in [3] under the same assumptions we have used. From it, the formula (D7) follows trivially.

Remark: It should be noted that the processes ν^i in (D4) are *not* innovations with respect to (Z_t,P) though they *are* conditional innovations processes given hypothesis h_i is true, i.e. with respect to $(Z_t \cap H_i,P)$. In other words, ν_i is not a (Z_t,P)-Wiener process (or martingale), but a $(Z_t \cap H_i,P)$-Wiener process. The innovations process with respect to (Z_t,P) is $d\bar{\nu}_t \triangleq dz_t - \bar{g}_t dt$, i.e. $\bar{\nu}_t$ is the conditional average of the $\nu_t^i, i=0,1,..,M-1$. Notice that $d\bar{\nu}_t$ appears, as one would expect, in (D3) and (D6), neither of which seem to have been written down before.

Remark: Equations (D3) and (D6) can be applied to combined detection and estimation problems, generalizing what has been done in [9] for doubly-stochastic observations (W), and to combined detection, estimation and control problems, extending the material in [10] which is restricted to linear state equations and linear observations in (W). Further developments along these lines will be reported elsewhere.

IV. REPRESENTATION THEOREMS

In order to derive a representation for the conditional distribution of x_s given $\mathcal{Z}_t$, we set $\Theta = I_A$ in (RCE3), where I_A is the indicator function of the set $A = \{\omega : x_s(\omega) \in B\}$ and $B \in \mathcal{B}_x$. Let μ_s, μ_s^*, $\mu_s(\cdot|\mathcal{Z}_t)$, and $\mu_s^*(\cdot|\mathcal{Z}_t)$ be the measures induced on $(S_x, \mathcal{B}_x)$ by P, P^*, $P(\cdot|\mathcal{Z}_t)$, and $P^*(\cdot|\mathcal{Z}_t)$, respectively.† Then it is a straightforward matter to show that (RCE3) becomes††

$$d\mu_s(x|\mathcal{Z}_t) = \left\{ E^*[L_t|\mathcal{Z}_t, x_s=x] \middle/ E^*[L_t|\mathcal{Z}_t] \right\} d\mu_s^*(x|\mathcal{Z}_t) \qquad \text{a.s.} \qquad (R1)$$

In order to reduce this expression to a more familiar and useful form we wish to

i) Convert $\mu_s^*(x|\mathcal{Z}_t)$ to $\mu_s(x)$

ii) Replace the expressions involving E^* with ones involving E.

To accomplish both, we require additional structure. Here, we assume that Z is doubly-stochastic with respect to X, so that from Corollary G we have X^* completely independent of Z^* and $P^*_X = P_X$. From this we have immediately $\mu_s^*(x|\mathcal{Z}_t) = \mu_s^*(x) = \mu_s(x)$, and (i) above is completed. Now fix attention to an element $z_t \in \mathcal{Z}_t$ corresponding to a fixed trajectory of z_s over $[0,t)$, so that, from (GW2) and (GC2), L_t becomes a function of X only, and thus $E^*[L_t|z_t] = E_X^*[L_t|z_t] = E_X[L_t|z_t]$, where the last equality follows from $P_X^*(\cdot|\mathcal{Z}_t) = P_X(\cdot|\mathcal{Z}_t)$. A similar argument holds for the numerator in (R1), and we have:

Representation 1: Let the process Z be given by either *doubly-stochastic* observation model (W) or (C). Then the conditional distribution $\mu_s(x|\mathcal{Z}_t)$ of x_s given $\mathcal{Z}_t$ has the representation

$$d\mu_s(x|\mathcal{Z}_t) = \left\{ E_X[L_t|\mathcal{Z}_t, x_s=x] \middle/ E_X[L_t|\mathcal{Z}_t] \right\} d\mu_s(x) \qquad \text{a.s.}$$

In particular, if the densities exist then

$$p_s(x|\mathcal{Z}_t) = \left\{ E_X[L_t|\mathcal{Z}_t, x_s=x] \middle/ E_X[L_t|\mathcal{Z}_t] \right\} p_s(x) \qquad \text{a.s.}$$

The following are trivial extensions:

Representation 2: In addition to the conditions of Representation 1, let $\{s_i; i \in I\}$ be any finite set of points in $[0,T]$. Then the joint conditional distribution $\mu_I(x_i; i \in I|\mathcal{Z}_t)$ of $\{x_{si}; i \in I\}$ given $\mathcal{Z}_t$ has the representation

$$d\mu_I(x_i; i \in I|\mathcal{Z}_t) = \left\{ E_X[L_t/\mathcal{Z}_t, x_{si}=x_i;\ i \in I] \middle/ E_X[L_t|\mathcal{Z}_t] \right\} d\mu_I(x_i; i \in I) \qquad \text{a.s.}$$

Representation 3: Let X be replaced by the combined process (X, Θ) in Representation 1. Then the conditional distribution of $\boldsymbol{\theta}_s$ given in $\mathcal{Z}_t$ is given by

$$d\nu_s(\boldsymbol{\theta}|\mathcal{Z}_t) = \left\{ E_{\Theta}[L_t|\mathcal{Z}_t,\ \boldsymbol{\theta}_s = \boldsymbol{\theta}] \middle/ E_{\Theta}[L_t|\mathcal{Z}_t] \right\} d\nu_s(\boldsymbol{\theta}) \qquad \text{a.s.} \qquad (*)$$

If, for each s, $\boldsymbol{\theta}$ takes only the values $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ with probabilities $P_s(\boldsymbol{\theta}_1), \ldots, P_s(\boldsymbol{\theta}_n)$, respectively, then

$$P_s(\boldsymbol{\theta}_i|\mathcal{Z}_t) = \left\{ E_{\Theta}[L_t|\mathcal{Z}_t,\ \boldsymbol{\theta} = \boldsymbol{\theta}_i] \middle/ E_{\Theta}[L_t|\mathcal{Z}_t] \right\} P_s(\boldsymbol{\theta}_i) \qquad \text{a.s.} \qquad (**)$$

Remark The above representations hold:

a) for both observation models (W) and (C)

b) for filtering (t=s), smoothing (t > s), *and* prediction (t < s)

---

†More correctly, the conditioning for $\mu_s(\cdot|\cdot)$ and $\mu_s^*(\cdot|\cdot)$ should be on the sub-σ-algebra of $\mathcal{B}_x$ that is equivalent to $\mathcal{Z}_t$ in the sense that they are related via $x_s(\cdot)$. No confusion should arise from this (standard) notational license.

††$E[L_t|\mathcal{Z}_t, x_s=x]$ is a suggestive (and standard) shorthand for "$E\ [L_t|\mathcal{Z}_t \vee \sigma(x_s)]$ evaluated at $x_s=x$".

c) without requiring X is an Ito process or the existence of densities.

They thus provide strict generalizations of all existing representation theorems for the conditional distribution or the conditional density. This appears to be the first time both observation models have been treated together in a unified development. The works that are closest in spirit and generality to this approach are those of Wong [11] for the additive Wiener process model and Bremaud [3] for the counting process model; both, however, consider only the filtering case. Representation 2 gives a generalization to distributions and to the observation model (C) of the representation in [12] for the joint conditional density, and thus it also generalizes [13]. The earlier representation theorems in [14]-[18] are more restricted than those discussed above. Representation 3(*) generalizes [19] in all three respects a) - c), while (**) has no direct counterpart in the literature but was suggested by a problem examined in a more restricted setting in [20]. Several authors have examined related problems involving counting processes and, more generally, independent increment processes, e.g. [8], [18], [21], [22]. Likelihood functions are known for a number of these formulations, but not analogs of the Girsanov Theorem GC. In particular, conditions for the equivalence of P and P^* are lacking, as is a characterization of the process Z^*.

Remark The principal restriction in all three representations is the requirement, common to all existing representation theorems, that Z be doubly-stochastic with respect to X. Its removal is an important open question. From the viewpoint of our development, it provides a sufficient condition for X^* and Z^* to be completely independent which, in turn, enables us to equate the conditional measure $\mu_s^*(x|\mathcal{Z}_t)$ induced by P^* to the unconditional measure $\mu_s(x)$ induced by P and, furthermore, plays an important role in converting E^* to E_X. For this latter step we must also restrict attention to elements of $\mathcal{Z}_t$ that correspond to a single observation trajectory over $[0,t]$, and this imposes a caveat on the interpretation of all three above representations and their predecessors: contrary to what the notation might indicate, these representations do not hold for arbitrary $z_t \in \mathcal{Z}_t$, though this is no restriction for applications.

It will be shown in later sections that the assumption that Z be doubly-stochastic with respect to X can be removed when (RCE3) is used to derive expressions for entities other than the conditional distribution, such as $E[h(x_s)|\mathcal{Z}_t]$ or the conditional probabilities associated with hypothesis testing.

Remark When X is a diffusion process, it is a simple and straightforward matter to use the Ito differential rule to derive stochastic differential equations for the various conditional densities and conditional probabilities represented above, and from these to derive stochastic differential equations for such entities as the conditional mean. In many cases these derivations are well-known or the stochastic differential equations have been derived by other means, though an exhaustive "search" turns up some extensions and equations that don't appear to have been written down explicitly in the literature. We do not pursue this here. In any case, many of these results follow more easily from the approach in Section V.

V. ESTIMATION

Let us consider $\Theta = h(x_s)$, $s \in [0,T]$, in (RCE3), where h is a $\sigma(x_s)$-measurable function and satisfies all required technical properties. Now define

$$\hat{h}_{s|t} \triangleq E[h(x_s)|\mathcal{Z}t] = E^*[h(x_s)L_t|\mathcal{Z}_t] \Big/ E^*[L_t|\mathcal{Z}_t] \qquad \text{(E1)}$$

which we rewrite for convenience as

$$\hat{h}_{s|t} = A_{st}\big/B_t, A_{st} = E^*[h(x_s)L_t|Z_t] \; , \; B_t = E^*[L_t|Z_t] \qquad \text{(E1)}'$$

We also assume, that X satisfies the Ito equation

$$dx_t = f(x_t,t)\,dt + b(x_t,t)\,dv_t \qquad \text{(E2)}$$

where f and b are such that the solution is well-defined, $V = \{v_t, F_t, P\}$ is a Wiener process, and the distribution of x_0 is known. For the observation model (W) we assume we assume that for all t the future dynamics noise $\{v_s, F_s, P;\ s \in [t,T]\}$ is independent of the past observation noise $\{w_s, F_s, P;\ s \in [0,t]\}$ and that x_0 is independent of w; this implies that Z is doubly-stochastic with respect to X. This assumption will be relaxed later. Whenever in the sequel (E2) and (W) are taken to be linear, we also assume x_0 to be Gaussian and independent of V. For the observation model (C), V and x_0 are assumed to be such that Z is doubly-stochastic with respect to X.

The derivation of a stochastic differential or integral equation for $\hat{h}_{s|t}$ is simply a matter of applying the appropriate Ito differential rule to (E1):

Wiener process models: Suppose A_{st} (for fixed s) and B_t satisfy stochastic d.e.'s of the form $d\zeta_t = \alpha_t dt + \beta_t dw_t$, where W is a Wiener process. Then

$$d\left(\frac{A_{st}}{B_t}\right) = \left(\frac{dA_{st}}{B_t}\right) - \left(\frac{A_{st}}{B_t}\right)\left(\frac{dB_t}{B_t}\right) - \left(\frac{dA_{st}}{B_t}\right)\left(\frac{dB_t}{B_t}\right)' + \left(\frac{A_{st}}{B_t}\right)\left(\frac{dB_t}{B_t}\right)\left(\frac{dB_t}{B_t}\right)' \qquad \text{(E3W)}$$

where we have used the well-known facts: $dw_t dt = 0$, $dw_t\, dw_t' = I\, dt$ [11].

Counting process models: Suppose A_{st} (for fixed s) and B_t satisfy stochastic d.e.'s of the form $d\zeta_t = \alpha_t dt + \beta_t dN_t$, where N is a counting process [3]. Then

$$d\left(\frac{A_{st}}{B_t}\right) = \left[\frac{dA_{st}}{B_t} - \frac{A_{st}}{B_t}\,\frac{dB_t}{B_t}\right]\left[1 + \sum_{n=1}^{\infty} (-1)^n \left(\frac{dB_t}{B_t}\right)^n\right] \qquad \text{(E3C)}$$

where we have used: $dN_t dt = 0$, $(dN_t)^n = dN_t$ for $n \geq 1$ [5].

We illustrate the technique by considering fixed-point smoothing ($t>s$, s constant) for the observations (C). Here application of a lemma in [23, Eqs. (18), (21)] and the Ito rule in [5] to A_{st}, B_t defined in (E1)' yields, after some simple algebra,

$$dA_{st} = E^*[h_x(g_t - 1)L_t|Z_t]\ (dz_t - dt)$$

$$dB_t = E^*[(g_t - 1)L_t|Z_t]\ (dz_t - dt)$$

where $h_s \triangleq h(x_s)$. Combining these with (E3C) and (E1)', collecting terms and integrating, gives

$$\hat{h}_{s|t} = \hat{h}_s + \int_s^t [\widehat{(h_s g_\sigma)}_\sigma - \hat{h}_{s|\sigma}\hat{g}_\sigma]\hat{g}_\sigma^{-1}[dz_\sigma - \hat{g}_\sigma d\sigma] \qquad \text{(E4)}$$

where $\hat{g}_t \triangleq E[g_t|Z_t]$, $\widehat{(h_s g_t)}_t \triangleq E[h_s g_t|Z_t]$.

the analogous expression for the observations (W) is just as easily found and differs from (E4) only in that $\hat{g}_t^{-1}$ is replaced by $W^{-1}(t)$ [5], [24]. In this case the stochastic d.e.'s for fixed-interval and fixed-lag smoothing follow by differentiation, exactly as in [25].

Equally simple derivations yield stochastic d.e.'s for both observation models for filtering and fixed-point prediction ($t<s=\text{constant}$), from which fixed-lead and fixed-internal prediction can be recovered by integration and differentiation. The resulting equations are known or can be deduced from equations for the conditional distribution or density, or by using the $\hat{g}_t^{-1} \leftrightarrow W^{-1}(t)$ correspondence between results for (C) and (W).

Rarely, however, are the existing derivations as short and direct as via the approach given here.

Extentions: 1. For fixed point smoothing, the assumption that X is given by an Ito process (E2) can be relaxed since no use is made of this: Z should still be doubly-stochastic with respect to X, however.

2. In filtering and smoothing problems, it is easy to generalize (E2) to replace the first term on the right side by $f(x_t,u_t,t)$ where u_t is Z_t-measurable and the usual technical assumptions are met, so paving the way for applications to control.

3. It is also simple to extend the filtering equations to allow for V in (E2) and W in (W) to be correlated. If $b(x_t,t) = D(t)$ in (E2), there is no loss of generality in writing

$$dv_t = \Gamma(t)G^{-1}(t)dw_t + d\bar{v}_t \,, \quad E[v_t w_s'] = \int_0^{t\wedge s} \Gamma(\tau)d\tau \tag{E5}$$

where $\bar{V}$ and W are completely independent, i.e. $E[d\bar{v}_t dw_s'] = 0$. Substituting $dw_t = G^{-1}(t)[dz_t - g(x_t,t)dt]$ into (E5) and the result into (E2) gives an Ito equation for x in terms of z_t and $\bar{v}_t$, which is used in a straightforward application of the Ito rule to (E1)' to yield

$$d\hat{x}_t = \hat{f}_t dt + [\widehat{(x_t g_t)}_t - \hat{x}_t \hat{g}_t' + D(t)\Gamma(t)]W^{-1}(t)\ [dz_t - \hat{g}_t dt]$$

Of course, this solution now paves the way for consideration of problems with "colored" noise.

VI. CONCLUDING REMARKS

The simple notion of relative conditional expectation provides the means for a novel utilization of measure transformation techniques in an attractive and powerful alternative approach to a wide variety of problems involving stochastic processes. The applications we have sketched give just an indication of the generality, versatility, and simplicity of this viewpoint. Further applications, for example to stochastic control problems, will be reported elsewhere.

VII. REFERENCES

[1] I.V. Girsanov, "On Transforming a Certain Class of Stochastic Processes by Absolutely Continuous Substitution of Measures", Theory of Prob. and Appl., 5, 1960, pp. 285-301.

[2] M. Loève, Probability Theory, 3rd ed., Van Nostrand, Princeton, N.J., 1963.

[3] P. Brémaud, "A Martingale Approach to Point Processes", Mem. ERL-M345, Electronics Research Lab., Dept. of EECS, U.C. Berkeley, 1972.

[4] T. Kailath, "A Further Note on a General Likelihood Formula for Random Signals in Gaussian Noise," IEEE Trans. on Info. Th., IT-16, 4, 1970, pp. 393-396.

[5] D.L. Snyder, "Filtering and Detection for Doubly-Stochastic Poisson Processes," IEEE Trans. on Info. Th., IT-18, 1, 1972, pp. 91-102.

[6] Yu. Sosoulin, R.L. Stratonovich, "Optimum Detection on a Diffusion Process in White Noise," Radio Engrg. Electron. Phys., 10, 1965, pp. 704-714.

[7] D.L. Snyder, "A Separation Theorem for the Detection of Doubly-Stochastic Poisson Process," Proc. of the 5th Ann. Princeton Conf. on Information Science and Systems, 1971, p. 286.

[8] P. Frost, "Notes on Conditionally Independent Increment Processes," lecture notes for *Current Trends in Automatic Control*, Washington University, July-August 1971.

[9] A.H. Haddad, J.B. Cruz, Jr., "Nonlinear Filtering for Systems with Random Models," Proc. 2nd Symp. on Nonlinear Est. Th. & Its Appl., San Diego, Calif., 1971, pp. 147-150.

[10] D. Willner, "Observation and Control of Partially Unknown Systems," Report No. ESL-R-496, Elec. Syst. Lab., Dept. of EE, M.I.T., Mass., 1973.

[11] E. Wong, Stochastic Processes in Information and Dynamical Systems, McGraw-Hill, N.Y., 1971.

[12] J.T. Lo, "On Optimal Nonlinear Estimation-Part I: Continuous Observation", Information Sciences, 6, 1973, pp. 19-32.

[13] B.B. Sahay, "Estimation in Continuous Time Systems Via New Representations," D.Sc. dissertation, CSSE, Washington Univ., St. Louis, Mo., 1973.

[14] R.S. Bucy, P.D. Joseph, Filtering for Stochastic Processes with Applications to Guidance, Interscience Publishers, Wiley, N.Y. 1968.

[15] R.E. Mortensen, "Optimal Control of Continuous-Time Stochastic Systems," Ph.D. dissertation, Rept. No. ERL-66-1, Univ. of Calif., Berkeley, 1966.

[16] T.E. Duncan, "Probability Densities for Diffusion Processes," TR No. 7001-4, Stanford Electronics Lab., Stanford Univ., Stanford, Calif., 1967.

[17] G. Kallianpur, C. Striebel, "Estimation of Stochastic Systems: Arbitrary System Process with Additive White Noise Observation Errors," Ann. Math. Statist., 39, 1968, pp. 785-801.

[18] D.L. Snyder, "A Representation Theorem for Observed Jump Processes," Proc. 1972 IEEE Conference on Decision & Control, December 1972, p. 218.

[19] D.G. Lainiotis, "Optimal Nonlinear Estimation," Int. J. Control, Vol. 14, 6, 1971, pp. 1137-1148.

[20] A.N. Shiryaev, "Stochastic Equations of Nonlinear Filtering of Markovian Jump Processes," Problems of Information Transmission, Vol. 2, 3, 1966, pp. 1-18.

[21] A. Segall, T. Kailath, "Radon-Nikodym Derivatives with Respect to Measures Induced by Discontinuous Independent-Increment Processes," Internal Report, Dept. of EE, Stanford Univer., Stanford, Calif., 1973.

[22] A.V. Skorokhod, Studies in the Theory of Radom Processes, Addison-Wesley, 1965.

[23] H.J. Kushnev, "Dynamical Equations for Optimal Nonlinear Filtering," J. of Diff. Eqns., 3, 1967, pp. 171-190.

[24] P. Frost, T. Kailath, "An Innovations Approach to Least-Squares Estimation-Part III: Nonlinear Estimation in White Gaussian Noise," IEEE Trans. on Automatic Control, AC-16, 3, 1971, pp. 217-226.

[25] T. Kailath, P. Frost, "An Innovations Approach to Least-Squares Estimation-Part II: Linear Smoothing in Additive White Noise," IEEE Trans. on Automatic Control, AC-13, 6, 1968, pp. 655-660.

STALEMATE AND OTHER PROBLEMS OF CONTROLLERS IN CONFLICT

DAHLARD L. LUKES
Department of Applied Mathematics and Computer Science
University of Virginia
Charlottesville, Virginia 22901

ABSTRACT

The article treats a conflicting end state problem for the linear differential game $\dot{x} = Ax + B_1u_1 + B_2u_2$ where the controllers are obtained by consideration of associated control problems for equations of the type $\dot{x} = Ax + Bu + f(t)$. The controllers of compensator type are singular integral operators. The problem of studying the behavior of $x(t)$ as $t \to T$ is dealt with by examining a related linear ordinary differential equation with singular coefficients.

INTRODUCTION

Underlying Control Problems

There are points in the development of the controllability theory for a differential equation at which the results can have a bearing upon certain questions of interest in differential games. The problem studied in this article illustrates how this can happen.

Consider the linear system in R^a

$$\dot{x}(t) = Ax(t) + Bu + f(t) \tag{1.1}$$

operating on $0 \leq t \leq T$ with boundary conditions

$$x(0) = x_0, \qquad x(T) = x_T \tag{1.2}$$

in which A, B are real matrices and u in R^b is the control variable. Assume that rank $[B, AB, \ldots, A^{a-1}B] = a$. A few pertinent facts about the controllability problem (1.1)-(1.2) are now summarized.

If $f(t)$ is identically zero then it is quite easy to show that there exists a unique linear controller of the form

$$u(t) = K(t)x(t) + r(t) \tag{1.3}$$

which automatically produces control signals solving (1.1)-(1.2). The matrix function $K(t)$ and vector $r(t)$ are restricted to depend upon only the data of the problem - A, B, x_0, x_T and T.

If f(t) is not identically zero then a similar remark applies but with the difference that now r(t) depends upon $f(\tau)$, $t \leq \tau \leq T$. In control problems involving unpredictable disturbances, f, this future information is not available. To cope with the difficulty a special controller of the type

$$u(t) = \int_0^t K(t,\omega) f(\omega) d\omega + r(t) \tag{1.4}$$

referred to by the author as a "compensator" was introduced in [1, 2]. Again K(t, ω) and r(t) depend upon only the data A, B, x_0, x_T and T. Precise formulas will be given below. It is clear that no future information about f is required in the generation of u(t).

In the cited references the K and r of (1.4) were selected so that (1.1), (1.2) and (1.4) are formally satisfied for every f(t). Due to a singularity occurring in K(t, ω) the compensator actually succeeds in doing the required steering of (1.1) for only a restricted class of f's. If the second boundary condition in (1.2), $x(T) = x_T$, is replaced by $\lim_{t \to T} x(t) = x_T$ (referred to as "asymptotic controllability") then the subclass can be expanded a bit and substantially so if one is willing to tolerate the possibility that $|u(t)| \to \infty$ as $t \to T$.

The reader can find more details and a treatment of nonlinear systems in [1, 2]. However, the above background information is all that is essential for understanding the present article.

Transition to Games

Our control theory tells us that for each linear equation in $R^a x[0, T]$

$$\dot{x}(t) = Ax(t) + Bu + f(t) \tag{2.1}$$

there exists a compensator

$$u(t) = \int_0^t K(t,\omega) f(\omega) d\omega + r(t) \tag{2.2}$$

which will automatically solve the steering problem

$$x(0) = x_0, \qquad \lim_{t \to T} x(t) = x_T \tag{2.3}$$

for a large class of f's. (We do not worry about how the class depends upon the data since in the physical interpretation of the problem we will not know which f's to expect anyway.) The condition that rank $[B, AB, \ldots, A^{a-1}B] = a$ is assumed.

Consider now the differential game equation

$$\dot{x}(t) = Ax(t) + B_1u_1 + B_2u_2 \tag{2.4}$$

in which player P_i commands control (vector) variable u_i and wishes to steer according to

$$x(0) = x_0, \qquad \lim_{t \to T} x(t) = x_i \qquad (i = 1, 2). \tag{2.5}$$

If $x_1 \neq x_2$ then there exists a conflict of interest. With no prior knowledge about what kind of steering P_2 might be expected to do it would be quite natural for P_1 to regard (2.4) as an equation of type (2.1) with $B_1 = B$ and $f(t) = B_2u_2(t)$. Hence he might employ a corresponding compensator of type (2.2) which would guarantee his success against a large class of control actions $u_2(t)$ taken by P_2. The same argument can be made by switching the roles of P_1 and P_2. This leads to precisely the equations studied in the next section.

A natural question to consider then is "what happens to $x(t)$ as $t \to T$?" The examples presented and the sufficiency conditions derived show that one possibility is that $x(t) \to \frac{x_1 + x_2}{2}$ as $t \to T$, i.e. a stalemate develops. A complete answer is not yet in hand. Resolution of the stalemate problem alone would be enough to open up possibilities for cooperative play in multi-player games.

A similar approach can carry one from the controllability of differential equations with nonlinear dynamics to problems in nonlinear games. Recent results in global control of nonlinear differential equations indicate that progress in this direction is more than a remote possibility of the future. (See [3, 4] for example.) The present article is intended to indicate the kind of mathematical problems one is led into when the underlying control system is linear.

A Linear Two-Player Game

Suppose that we consider the game equation in $R^a x [0, T]$

$$\dot{x} = Ax + B_1u_1 + B_2u_2, \qquad x(0) = x_0 \,. \tag{3.1}$$

Assume player P_i commands variable u_i. If each player regards the other as a disturbance we are led to the pair of compensators derived in [1]

$$u_i(t) = r_i(t) + \int_0^t K_i(t, \omega)\, B_j\, u_j(\omega)\, d\omega \qquad j \neq i;\ i, j = 1, 2 \tag{3.2}$$

where

$$r_i(t) = B_i^* e^{A^*(T-t)} S_i^{-1}(T)\ (x_i - e^{AT} x_0) \tag{3.3}$$

$$K_i(t, \omega) = -B_i^* e^{A^*(T-t)} S_i^{-1}(T - \omega)\ e^{A(T-\omega)} \tag{3.4}$$

$$S_i(\varepsilon) = \int_0^{\varepsilon} e^{A\sigma} B_i B_i^* e^{A^*\sigma} d\sigma, \ (i = 1, 2) \tag{3.5}$$

and x_i denotes the state to which player P_i wishes to steer the system. We investigate the behavior of $x(t)$ as $t \to T$.

Note that the equations (3.2) are a pair of linear coupled Volterra equations which will have a unique solution on $0 \leq t < T$ which can be substituted into (3.1) to generate the corresponding response. There is no problem with invertibility of S_i in formulas (3.3)-(3.4) because, denoting rank $[B, AB, \ldots, A^{a-1}B] = k(A, B)$, we will always assume $k(A, B_i) = a$, $(i = 1, 2)$. This condition is equivalent to invertibility of $S_i(\varepsilon)$ for $\varepsilon \neq 0$ and in turn equivalent to the controllability of (3.1) relative to u_i.

Example 3.1 $\qquad \dot{x} = u_1 + u_2$

Using the two compensators

$$u_1(t) = \frac{x_1 - x_0}{T} - \int_0^t \frac{u_2(\omega)}{T - \omega} d\omega$$

$$u_2(t) = \frac{x_2 - x_0}{T} - \int_0^t \frac{u_1(\omega)}{T - \omega} d\omega$$

we compute out the responses

$$u_1(t) = \frac{x_1 - x_2}{2(T - t)} + \left[\frac{x_1 + x_2 - 2x_0}{2T^2}\right](T - t)$$

$$u_2(t) = \frac{x_2 - x_1}{2(T - t)} + \left[\frac{x_1 + x_2 - 2x_0}{2T^2}\right](T - t)$$

$$x(t) = x_0 + \left[\frac{x_1 + x_2 - 2x_0}{T^2}\right](T - \frac{t}{2})t \;.$$

Note the pole in $u_i(t)$ due to the conflicting goals. The operators will always produce a stalemate. Moreover if the game is started at the stalemate state it will remain there. Before proceeding with the general stalemate question it is interesting to note what happens in the next example where we feed back the updated state as well.

Example 3.2 $\dot{x} = u_1 + u_2$

The feedback operators are then

$$u_1(t) = \frac{x_1 - x}{T - t} - \int_0^t \frac{u_2(\omega)}{T - \omega}\,d\omega$$

$$u_2(t) = \frac{x_2 - x}{T - t} - \int_0^t \frac{u_1(\omega)}{T - \omega}\,d\omega \;.$$

Reduce the differential equation to

$$\ddot{x} + \frac{3}{T - t}\dot{x} + \frac{2}{(T - t)^2}x = \frac{x_1 + x_2}{(T - t)^2} \;.$$

Try $x = (T - t)^r$. Therefore $r = 2 \pm \sqrt{2}$ and one can compute

$$x(t) = \frac{1}{2T^2}\left[x_0 - \frac{(x_1 + x_2)}{2}\right]\left[\frac{(T - t)^{2+\sqrt{2}}}{T^{\sqrt{2}}} + \frac{(T - t)^{2-\sqrt{2}}}{T^{-\sqrt{2}}}\right] + \frac{x_1 + x_2}{2} \;.$$

Clearly a stalemate again occurs. Note that $x(t)$ is continuous at $t = T$ but not differentiable there. $|\dot{x}(t)| \to \infty$ as $t \to T$ showing the controls are going unbounded. We now return to the study of the limiting behavior of the solution $x(t)$ to (3.1)-(3.2).

It is convenient to introduce the following linear differential equation with singular coefficients

$$\dot{\rho}(t) = [R_1(t) + R_2(t)]\rho(t) + [R_1(t) - R_2(t)](x_2 - x_1) \qquad (3.6)$$

$$\rho(0) = 2\,e^{AT}x_0 - (x_1 + x_2) \qquad (3.7)$$

where

$$R_i(t) = \frac{dS_i(T - t)}{dt} S_i^{-1}(T - t). \qquad (3.8)$$

Recall that S_i is defined by (3.5).

The next theorem summarizes our conversion of the problem to one of ordinary differential equations. Details of the proof are given in [2].

Theorem 3.1 *The solution* $x(t)$ *of* (3.1)-(3.2) *with* $k(A, B_i) = a$ $(i = 1, 2)$ *has a limit as* $t \uparrow T$ *if and only if the solution* $\rho(t)$ *to* (3.6)-(3.7) *does and moreover the limits are related by*

$$\lim_{t\to T} x(t) = \frac{x_1 + x_2}{2} + \frac{1}{2} \lim_{t\to T} \rho(t) \ . \tag{3.9}$$

We would like to examine the behavior of $\rho(t)$ as $t \uparrow T$. The next lemma will be useful for examining the nature of the singularities in the coefficients of (3.6).

Lemma 3.1 *The series expansion of the integral* (3.5)

$$S(\varepsilon) = \sum_{i=0}^{\infty} \frac{\varepsilon^{i+1}}{i+1} \sum_{j=0}^{i} \frac{1}{j!(i-j)!} A^j BB^*A^{*(i-j)} \tag{3.10}$$

is valid for all ε. *Under the controllability assumption that* $k(A, B) = a$ *there exists a smallest integer* $N \geq 0$ *such that*

$$\hat{S}(\varepsilon) \overset{df}{=} \sum_{i=0}^{N} \frac{\varepsilon^{i+1}}{i+1} \sum_{j=0}^{i} \frac{1}{j!(i-j)!} A^j BB^*A^{*(i-j)} \tag{3.11}$$

is positive definite for all $\varepsilon > 0$ *near zero.*

$$\left[\hat{S}(\varepsilon)\right]^{-1} = \frac{P(\varepsilon)}{\varepsilon^p} \tag{3.12}$$

where the matrix $P(\varepsilon)$ *of real rational functions of* ε *is selected so that the integer* $p \geq 1$ *is minimal.*

Proof. The expansion (3.10) can be rigorously derived by replacing the exponential functions of the integrand in (3.5) by series and reversing the order of integration and summation. Notice that every truncation of that series is a symmetric matrix function of ε. The proof is by contradiction. See [2] for the details. The following stability-theoretic result was proved by the author in [2] using a Lyapunov argument.

Lemma 3.2 *Consider the differential equation in* $R^a \times [0,\infty)$

$$\frac{dy}{dt} = A(t)y + f(t) \tag{3.13}$$

where $A(t)$, $f(t)$ *are continuous* $A(t) \to A_\infty$ *and* $f(t) \to 0$ *as* $t \to \infty$ *with* $\sigma(A_\infty) \subseteq [z \mid \mathrm{Re}\ z < 0]$. *Then* $y(t) \to 0$ *as* $t \to \infty$ *for each solution* $y(t)$ *of* (3.13).

Corollary 3.1 *If the matrices* B_i $(i = 1, 2)$ *in* (3.1) *are nonsingular then a stalemate will always develop* (i.e. $x(t) \to \frac{x_1 + x_2}{2}$ as $t \to T$) *whenever the players employ the respective strategies given by* (3.2) - (3.5).

Proof. First note that with the B_i nonsingular the condition $k(A, B_i) = a$ required to make (3.5) nonsingular is automatically met.

The proof is based upon Theorem 3.1 and Lemma 3.1. With the B_i nonsingular we need compute only the first term in Lemma 3.1 to get $N_i = 0$, $\hat{S}_i(\varepsilon) = \varepsilon B_i B_i^*$, $P_i(\varepsilon) = I$ and $p_i = 1$, $(i = 1, 2)$. Since

$$S_i(\varepsilon) = \hat{S}_i(\varepsilon) + 0_i(\varepsilon^2) \tag{3.14}$$

where $|0_i(\varepsilon^2)| \leq c\,\varepsilon^2$ for some constant c we can estimate

$$S_i^{-1}(\varepsilon) = [I + \hat{S}_i^{-1}(\varepsilon)\ 0_i(\varepsilon^2)]^{-1}\ \hat{S}_i^{-1}\ (\varepsilon) \tag{3.15}$$

using the fact that $|\hat{S}_i^{-1}(\varepsilon)\ 0_i(\varepsilon^2)| = 0(\varepsilon)$ to get

$$S_i^{-1}(\varepsilon) = \hat{S}_i^{-1}(\varepsilon) + 0_i(1) \tag{3.16}$$

with $|0_i(1)|$ bounded for all $\varepsilon > 0$ sufficiently small. Computing the coefficients in (3.6) using (3.8) we get

$$\begin{aligned} R_i(t) &= e^{-A(T-t)} B_i B_i^* e^{A^*(T-t)} S_i^{-1}(T - t) \\ &= - B_i B_i^*\ [\hat{S}_i^{-1}(T - t) + 0_i(1)] + 0(1) \qquad (3.17) \\ &= - \frac{I}{T - t} + 0(1) \end{aligned}$$

where again $|0(1)|$ is bounded. Therefore we can write (3.6) as

$$\frac{d\rho}{dt} = [- \frac{2I}{T - t} + 0(1)]\rho + 0(1). \tag{3.18}$$

Then making the change of variable $T - t = e^{-\tau}$ so that $\frac{dt}{d\tau} = (T - t)$ we transform (3.18) into

$$\frac{d\rho}{d\tau} = [- 2I + 0(e^{-\tau})]\rho + 0(e^{-\tau}). \tag{3.19}$$

Since (3.19) satisfies the hypothesis of Lemma 3.2 we conclude that $\rho \to 0$ as $\tau \to \infty$, i.e. as $t \to T$. The proof is now completed by appealing to Theorem 3.1.

Remark

Although the stability-theoretic approached taken in proving Corollary 3.1 worked well many systems cannot be treated in this manner for the reason that the limit matrix A_∞ associated with (3.6) fails to have all its eigenvalues in the open half-plane as required. More refined techniques, e.g. methods of asymptotic expansions, might be useful in determining the qualitative nature of the solution $\rho(t)$ as $t \to T$ in these other cases.

Some results, such as those in the next corollary, can be obtained from direct examination of (3.6) - (3.7).

Corollary 3.2 *If in the game* (3.1) *with* $k(A, B_i) = a$ $(i = 1, 2)$ *the players employ their respective strategies* (3.2) *then a stalemate will develop if*

(a) $\frac{x_1 + x_2}{2} = e^{AT}x_0$ *and either*

(b) $B_1B_1^* = \lambda B_2B_2^*$ *for some scalar* λ *or*

(c) $x_1 = x_2$.

Remarks about Corollary 3.2

Condition (a) centers the target points about the natural state the game would achieve at $t = T$ under zero control.

Property (b) is an expression of equi-potent placement of the players control variables in (3.1). In [5, 6] such games were found to always possess equilibrium strategies.

Under the concordance expressed by (c) the stalemate point makes everybody happy including old Mother Nature (the uncontrolled dynamics).

Proof of Corollary 3.2 One can easily check that if $\lambda \neq 0$ (which is the case since $k(A, B_i) = a$) then (b) implies that $R_1(t) = R_2(t)$ as computed by (3.8). Hence in case either (b) or (c) holds (3.6) is a homogeneous linear equation and (a) implies its solution satisfies $\rho(0) = 0$. By uniqueness we conclude $\rho(t)$ is identically zero. Theorem 3.1 then applies directly to conclude the proof.

REFERENCES

[1] D. L. Lukes, Global controllability of disturbed nonlinear equations, to appear in SIAM J. Control.

[2] _________, Asymptotic controllability of nonlinear differential games, to appear in the Proceedings of the Regional Research Conference on Differential Games and Control Theory held at the University of Rhode Island, June 1973.

[3] _________, Global controllability of nonlinear systems, SIAM J. Control, vol. 10, no. 1, February 1972 (Erratum: vol. 11, no. 1, February 1973).

[4] _________, Globally controllable differential equations with unbounded and bounded nonlinearities, to appear.

[5] _________, Equilibrium feedback control in linear games with quadratic costs, SIAM J. Control, vol. 9, no. 2, May 1971.

[6] D. L. Lukes and D. L. Russell, A global theory for linear quadratic differential games, J. Math. Anal. Appl., 33 (1971).

MINIMAX ADMISSIBLE ESTIMATORS OF THE MEAN OF A MULTIVARIATE NORMAL DISTRIBUTION WITH NON-IDENTITY COVARIANCE MATRIX

MARY ELLEN BOCK
University of Illinois
Urbana, Illinois

ABSTRACT

Given pxp positive definite matrices C and D, a real-valued function h and $p \geq 3$, let $X \sim N(\theta, D)$ and define $\hat{\theta}(X) = (I_p - Ch(X'D^{-1}X)X)$. Unless $h \equiv 0$ a.e., minimax estimators of this form exist if and only if $\mathrm{trace}(CD)/\xi_L > 2$ (where ξ_L is the largest eigenvalue of $D^{\frac{1}{2}}CD^{\frac{1}{2}}$). For $C=I_p$, the identity matrix, proper Bayes minimax estimators of the form $\hat{\theta}$ exist if and only if $\mathrm{trace}(CD)/\xi_L > \frac{p}{2} + 2$.

I. EVALUATION OF THE RISK FUNCTION

Assume the px1 random vector X has normal distribution with unknown mean θ and known positive definite covariance matrix D. Let $p \geq 3$ and define the risk function of an estimator $\hat{\theta}(X)$ of θ to be $R(\hat{\theta},\theta) = E(\hat{\theta}(X)-\theta)'(\hat{\theta}(X)-\theta)$. Consider an estimator of the form $\hat{\theta}(X) = (I_p - Ch(X'D^{-1}X))X$ where C is a pxp known positive definite matrix, where I_p is the pxp identity matrix and h is a real-valued measurable function. Then

$$R(\hat{\theta},\theta) = \mathrm{tr}D + E[h^2(X'D^{-1}X)X'C^2X] - 2E[h(X'D^{-1}X)X'CX] + 2E[h(X'D^{-1}X)X']C\theta. \qquad (1)$$

The following general theorems were developed to evaluate the terms on the right hand side.

Theorem 1: Let $Y \sim N(\eta, I_p)$ and $h:[0,\infty) \to (-\infty,+\infty)$. Then $E[h(Y'Y)Y_i] = \eta_i E h(\chi^2_{(p+2,\eta'\eta)})$.

Proof: Note that the Y_i's are elements of the px1 vector Y and they are independent. Therefore

$$E[h(Y'Y)Y_i] = E\{E[h(Y_i^2 + \sum_{j\neq i} Y_j^2)Y_i \Big| \sum_{j\neq i} Y_j^2]\} = E\left[e^{-\frac{\eta_i^2}{2}}\left(\int_{-\infty}^{+\infty} h(x^2 + \sum_{j\neq i} Y_j^2)\frac{xe^{\frac{-x^2}{2}}e^{x\eta_i}}{\sqrt{2\pi}}dx\right)\right]$$

$$= E\left[\frac{e^{-\frac{\eta_i^2}{2}}}{\sqrt{2\pi}}\left\{\int_0^\infty h(y + \sum_{j\neq i} Y_j^2)e^{\frac{-y}{2}}\left(e^{\eta_i y^{\frac{1}{2}}} - e^{-\eta_i y^{\frac{1}{2}}}\right)\frac{dy}{2}\right\}\right] \text{ (making the transformation } y=x^2\text{)}$$

$$= E\left[\frac{e^{-\frac{\eta_i^2}{2}}}{2\sqrt{2\pi}}\left\{\int_0^\infty h(y + \sum_{j\neq i} Y_j^2)e^{-\frac{y}{2}}\left(\sum_{k=0}^{\infty}\frac{2(\eta_i y^{\frac{1}{2}})^{2k+1}}{(2k+1)!}\right)dy\right\}\right]$$

$$= \eta_i E\left[\int_0^\infty h(y + \sum_{j\neq i} Y_j^2)e^{-\frac{\eta_i^2}{2}}\left(\sum_{k=0}^{\infty}\frac{\left(\frac{\eta_i^2}{2}\right)^k}{k!}\frac{y^{\frac{2k+3}{2}-1}e^{-\frac{y}{2}}}{\Gamma(\frac{2k+3}{2})2^{\frac{2k+3}{2}}}\right)dy\right] \text{ (because } \Gamma(2k) = \Gamma(k)\Gamma(k+\tfrac{1}{2})2^{2k-1}/\sqrt{\pi}\text{)}$$

$$= \eta_i E[h(\chi^2_{(3,\eta_i^2)} + \sum_{j\neq i} Y_j^2)] = \eta_i E[h(\chi^2_{(3,\eta_i^2)} + \chi^2_{(p-1, \sum_{j\neq i}\eta_j^2)})]$$

$$\eta_i E[h(\chi^2_{(p+2,\eta'\eta)})].$$

q.e.d.

Theorem 2; Let $X \sim N(\theta,D)$ where D is a pxp positive definite matrix and let $h:[0,\infty) \to (-\infty,+\infty)$. Then $E[h(X'D^{-1}X)X] = Eh(\chi^2_{(p+2,\theta'D^{-1}\theta)})\theta$.

Proof: Let $Y = D^{-\frac{1}{2}}X$ (where $D^{\frac{1}{2}}$ is a positive definite matrix such that $D^{\frac{1}{2}}D^{\frac{1}{2}} = D$). Then $Y \sim N(D^{-\frac{1}{2}}\theta, I_p)$ and

$E[h(X'D^{-1}X)X] = D^{\frac{1}{2}}E[h(Y'Y)Y] = D^{\frac{1}{2}}(D^{-\frac{1}{2}}\theta)E[h(\chi^2_{(p+2,(D^{-\frac{1}{2}}\theta)'D^{-\frac{1}{2}}\theta)})]$

(by Theorem 1) $= \theta E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})]$.

q.e.d.

Theorem 3: Let $Y \sim N(\eta, I_p)$ and $h:[0,\infty) \to (-\infty,+\infty)$. Then $E[h(Y'Y)Y_i^2]$ $= E[h(\chi^2_{(p+2,\eta'\eta)})] + \eta_i^2 E[h(\chi^2_{(p+4,\eta'\eta)})]$, where $\chi^2_{(j,\lambda)}$ has a chi-square distribution with j degrees of freedom and noncentrality parameter λ.

Proof: Recall that a noncentral chi-square random variable $\chi^2_{(j,\lambda)}$ may be represented as χ^2_{j+2K} where $K \sim \text{Poisson}(\frac{\lambda}{2})$ and the conditional distribution of χ^2_{j+2K} given $K = k$ is that of the central chi-square with j+2k degrees of freedom. Note that Y_i's are independent. Therefore,

$$E[h(Y'Y)Y_i^2] = E\{E[h(Y_i^2 + \sum_{j\neq i} Y_j^2)Y_i^2 \,|\, \sum_{j\neq i} Y_j^2]\} = E\{E[h(\chi^2_{1+2K} + \sum_{j\neq i} Y_j^2)\chi^2_{1+2K} \,|\, \sum_{j\neq i} Y_j^2]\}$$

$$E\left\{e^{-\frac{\eta_i^2}{2}} \sum_{k=0}^{\infty} \frac{\left(\frac{\eta_i^2}{2}\right)^k}{k!} E[h(\chi^2_{1+2k} + \sum_{j\neq i} Y_j^2)\chi^2_{1+2k} \,|\, \sum_{j\neq i} Y_j^2]\right\}$$

$$E\left\{e^{-\frac{\eta_i^2}{2}} \sum_{k=0}^{\infty} \frac{\left(\frac{\eta_i^2}{2}\right)^k}{k!} (1+2k)\, E[h(\chi^2_{(3+2k)} + \sum_{j\neq i} Y_j^2) \,|\, \sum_{j\neq i} Y_j^2]\right\}$$

$$= E[h(\chi^2_{3+2K} + \sum_{j\neq i} Y_j^2)] + \left\{e^{-\frac{\eta_i^2}{2}} \sum_{k=0}^{\infty} \frac{\left(\frac{\eta_i^2}{2}\right)^k}{k!} (2k) Eh(\chi_{3+2K} + \sum_{j\neq i} Y_j^2)\right\}$$

$$= E[h(\chi^2_{(3,\eta_i^2)} + \sum_{j\neq i} Y_j^2)] + \left\{e^{-\frac{\eta_i^2}{2}} \eta_i^2 \sum_{k=1}^{\infty} \frac{\left(\frac{\eta_i^2}{2}\right)^{k-1}}{(k-1)!} E[h(\chi^2_{5+2(k-1)} + \sum_{j\neq i} Y_j^2)]\right\}$$

$$= E[h(\chi^2_{(p+2,\eta'\eta)})] + (\eta_i^2)\, E[h(\chi^2_{(p+4,\eta'\eta)})], \quad i=1, \ . \ . \ ., p.$$

(because $\sum_{j\neq i} Y_j^2 \sim \chi^2_{(p-1,\, \sum_{j\neq i} \eta_j^2)}$ and because $\sum_{j\neq i} Y_j^2$ and Y_i^2 are independent.)

q.e.d.

Theorem 4: Assume D and W are pxp positive definite matrices. Let $X \sim N(\theta,D)$ and let $h:[0,\infty) \to (-\infty,+\infty)$. Then $E[h(X'D^{-1}X)X'WX] =$ $\text{tr}(WD)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})] + \theta'W\theta E[h(\chi^2_{(p+4,\theta'D^{-1}\theta)})]$.

Proof: Let $D^{\frac{1}{2}}$ be a positive definite matrix such that $D^{\frac{1}{2}}D^{\frac{1}{2}} = D$. Choose an orthogonal matrix P such that $P'D^{\frac{1}{2}}WD^{\frac{1}{2}}P = \begin{bmatrix}\gamma_1 & & 0 \\ & \ddots & \\ 0 & & \gamma_p\end{bmatrix}$ where the γ_i are the positive characteristic roots of the positive definite matrix $D^{\frac{1}{2}}WD^{\frac{1}{2}}$. Let $Z = P'D^{-\frac{1}{2}}X$. Then $Z \sim N(P'D^{-\frac{1}{2}}\theta, I_p)$ and

$$E[h(X'D^{-1}X)X'WX] = E[h((X'D^{-\frac{1}{2}}P)(P'D^{-\frac{1}{2}}X))X'D^{-\frac{1}{2}}P\{P'D^{\frac{1}{2}}WD^{\frac{1}{2}}P\}P'D^{-\frac{1}{2}}X] =$$

$$E\left[h(Z'Z)Z'\begin{bmatrix}\gamma_1 & & 0\\ & \ddots & \\ 0 & & \gamma_p\end{bmatrix}Z\right] = \sum_{i=1}^{p}\gamma_i E[h(Z'Z)Z_i^2] = \sum_{i=1}^{p}\gamma_i\{E[h(\chi^2_{(p+2,(P'D^{-\frac{1}{2}}\theta)'(P'D^{-\frac{1}{2}}\theta))})]$$

$$+ (P'D^{-\frac{1}{2}}\theta)_i^2 E[h(\chi^2_{(p+4,(P'D^{-\frac{1}{2}}\theta)'(P'D^{-\frac{1}{2}}\theta))})]\} = \left(\sum_{i=1}^{p}\gamma_i\right)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})] +$$

$$\left(\sum_{i=1}^{p}\gamma_i(P'D^{-\frac{1}{2}}\theta)_i^2\right)E[h(\chi^2_{(p+4,\theta'D^{-1}\theta)})] = \mathrm{tr}(WD)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})] +$$

$$(P'D^{-\frac{1}{2}}\theta)'\begin{bmatrix}\gamma_1 & & 0\\ & \ddots & \\ 0 & & \gamma_p\end{bmatrix}(P'D^{-\frac{1}{2}}\theta)E[h(\chi^2_{(p+4,\theta'D^{-1}\theta)})] \text{ (because } \sum_{i=1}^{p}\gamma_i = \mathrm{tr}(D^{\frac{1}{2}}WD^{\frac{1}{2}}) =$$

$$\mathrm{tr}(WD)) = \mathrm{tr}(WD)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})] + \theta'W\theta E[h(\chi^2_{(p+4,\theta'D^{-1}\theta)})].$$

q.e.d.

(Theorems 1,2,3, and 4 are given in a forthcoming Annals of Statistics paper [2]).

Applying Theorem 2 to the second and third terms of the right hand side of (1) and applying Theorem 4 to the last term, we have

$$R(\theta,\theta) = \mathrm{tr}D + \mathrm{tr}(C^2D)E[h^2(\chi^2_{(p+2,\theta'D^{-1}\theta)})] + \theta'C^2\theta E[h^2(\chi^2_{(p+4,\theta'D^{-1}\theta)})] - 2(\theta'C\theta)E[h(\chi^2_{(p+4,\theta'D^{-1}\theta)})] - 2\mathrm{tr}(CD)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})] + 2(\theta'C\theta)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})]. \tag{2}$$

II. A CERTAIN CLASS OF MINIMAX ESTIMATORS

Let $\hat{\theta}(X)$ be defined as in Section I and define $h(y) = c\frac{r(y)}{y}$ where $r:[0,\infty) \to [0,1]$ is a monotone nondecreasing function and c is a positive constant. Let $D^{\frac{1}{2}}$ be a positive, definite matrix such that $D^{\frac{1}{2}}D^{\frac{1}{2}} = D$. Let γ_L and ξ_L be the largest eigenvalues of $D^{\frac{1}{2}}C^2D^{\frac{1}{2}}$ and $D^{\frac{1}{2}}CD^{\frac{1}{2}}$ respectively.

Theorem 5: If $c \leq 2\gamma_L^{-1}(\mathrm{trace}(CD)-2\xi_L)$, then $\hat{\theta}$ is minimax.

Proof: It suffices to show $0 \geq R(\hat{\theta},\theta)-\mathrm{tr}D$. By (2), for $K \sim \mathrm{Poisson}\left(\frac{\theta'D^{-1}\theta}{2}\right)$,

$$R(\hat{\theta},\theta)-\mathrm{tr}D = \mathrm{tr}(C^2D)E\left[\frac{c^2r^2(\chi^2_{(p+2+2K)})}{(\chi^2_{(p+2+2K)})^2}\right] + (\theta'C^2\theta)E\left[\frac{c^2r^2(\chi^2_{(p+4+2K)})}{(\chi^2_{(p+4+2K)})^2}\right]$$

$$-2(\theta'C\theta)E\left[\frac{cr(\chi^2_{(p+4+2K)})}{\chi^2_{(p+4+2K)}}\right] - 2\mathrm{tr}(CD)E\left[\frac{cr(\chi^2_{(p+2+2K)})}{\chi^2_{(p+2+2K)}}\right] + 2(\theta'C\theta)E\left[\frac{cr(\chi^2_{(p+2+2K)})}{\chi^2_{(p+2+2K)}}\right]$$

$$= E\left[\frac{cr(\chi^2_{(p-2+2K)})}{(p-2+2K)(p+2K)}\left\{(p-2+2K-\chi^2_{(p-2+2K)})\left(\mathrm{tr}CD + \left(\frac{\theta'C\theta}{\theta'D^{-1}\theta}\right)2K\right) + \mathrm{tr}C^2D\left(cr(\chi^2_{(p-2+2K)})\right.\right.\right.$$

$$\left.\left.\left. - 2(p-2)\frac{\mathrm{tr}(CD)}{\mathrm{tr}(C^2D)} + \frac{\theta'C^2\theta}{\theta'D^{-1}\theta}2K\left(cr(\chi^2_{(p-2+2K)}) - 2\frac{\theta'D^{-1}\theta}{\theta'C^2\theta}\left(\mathrm{tr}CD - 2\frac{\theta'C\theta}{\theta'D^{-1}\theta}\right)\right)\right)\right\}\right]$$

(using the facts that $E\left[\frac{g(\chi^2_{j+2K})}{\chi^2_{j+2K}}\right] = E\left[\frac{g(\chi^2_{j-2+2K})}{(j-2+2K)}\right]$ and $E[2Kg(\chi^2_{j+2K})] = E[\theta'D^{-1}\theta g(\chi^2_{j-2+2K})]$ where g is a measurable function with finite expectation and $j \geq 3$.). Now $\left(\mathrm{tr}CD - 2\frac{\theta'C\theta}{\theta'D^{-1}\theta}\right)\frac{\theta'D^{-1}\theta}{\theta'C^2\theta} \geq (\mathrm{tr}CD-2\xi_L)\gamma_L^{-1}$ and (for all $t \geq 0$) $2\gamma_L^{-1}(\mathrm{tr}CD-2\xi_L) \geq cr(t)$ implies

$$R(\hat{\theta},\theta)-\mathrm{tr}D \leq E\left[\frac{cr(\chi^2_{(p-2+2K)})}{(p-2+2K)(p+2K)}\left\{2(p-2+2K-\chi^2_{(p-2+2K)})\left(\mathrm{tr}CD + \frac{\theta'C\theta}{\theta'D^{-1}\theta}\ 2K\right)\right.\right.$$
$$\left.\left.+\ \mathrm{tr}C^2D\left(2\gamma_L^{-1}(\mathrm{tr}CD-2\xi_L)-2(p-2)\frac{\mathrm{tr}CD}{\mathrm{tr}C\ D}\right)\right\}\right] \leq E\left[\frac{cr(\chi^2_{(p-2+2K)})}{(p-2+2K)(p+2K)}\ 2(p-2+2K-\chi^2_{(p-2+2K)})\left(\mathrm{tr}CD + \frac{\theta'C\theta}{\theta'D^{-1}\theta}\ 2K\right)\right]$$

(because $\mathrm{tr}CD \leq p\xi_L$ and $\mathrm{tr}C^2D \leq p\gamma_L$) < 0 (because $E[g(Y)(EY-Y)] \leq 0$ if Y is nonnegative and g is monotone nondecreasing).

q.e.d.

(This theorem has been shown by Baranchik [1] for $D=C=I_p$.)

Because a proper Bayes estimator is admissible, we may specify a distribution for θ and the resulting proper Bayes estimator will be admissible. Assume $C = I_p$ and $\frac{\mathrm{tr}CD}{\xi_L} > \frac{p}{2} + 2$. Then choosing a such that $0 \leq a < 1$ and $\frac{\mathrm{tr}(D)}{\xi_L} \geq \frac{p}{2} + 3-a$, let the conditional distribution of θ given λ be $N(0,D\frac{(1-\lambda)}{\lambda})$ where the unconditional density of λ is given by $\lambda^{-a}(1-a)$ for $0 < \lambda < 1$. The admissible proper Bayes estimator with respect to this prior distribution on θ is given by

$$\delta(X) = \left[1 - \left(p+2-2a-2e^{-\frac{X'D^{-1}X}{2}}\left\{\int_0^1 \lambda^{\frac{p}{2}-a}\, e^{\frac{\lambda}{2}(X'D^{-1}X)}\, d\lambda\right\}^{-1}\right)\Big/(X'D^{-1}X)\right] X$$

(using integration by parts). Let $c = p+2-2a$ and define $r(t) =$

$$1- \frac{2e^{-\frac{t}{2}}\left\{\int_0^1 \lambda^{\frac{p}{2}-a}\, e^{-\frac{\lambda t}{2}}\, d\lambda\right\}^{-1}}{(p+2-2a)} \quad \text{for } t \geq 0.$$

It has been shown by Strawderman [7], p. 387, that $r(\cdot)$ is nondecreasing and $0 \leq r(\cdot) \leq 1$. Applying Theorem 5 with $C = I_p$, we have $\delta(X)$ is minimax when $c = p+2-2a \leq 2\gamma_L^{-1}(\mathrm{tr}CD-2\xi_L)$. Thus, if $\frac{\mathrm{tr}CD}{\xi_L} \geq \frac{p}{2} + 3-a$, $\delta(X)$ is a minimax admissible estimator for θ where $0 \leq a < 1$. Thus a minimax admissible estimator for θ exists when $C = I_p$ and $\frac{\mathrm{tr}CD}{\xi_L} > \frac{p}{2} + 2$. (The above has been given by Strawderman [7] for the case $C=D=I_p$.)

III. AN EXISTENCE CONDITION FOR THE MINIMAX ESTIMATOR

In Section II, examples of minimax estimators of the form $\hat{\theta}(X) = (I_p-Ch(X'D^{-1}X))X$ were given under the condition that $\mathrm{tr}CD > 2\xi_L$. Let $\hat{\theta}_0(X) = X$. If we identify with $\hat{\theta}_0$ those estimators that coincide with $\hat{\theta}_0$ almost everywhere with respect to Lebesgue measure, then we may show that $\hat{\theta}_0$ is the only minimax estimator of the form $\hat{\theta}$. The following lemmas, 1 and 2, lead to this result in Theorem 6.

For $p=2$ and $C=D=I_2$ and $\delta(X)=g(X'X)X$, setting $h(\cdot)=1-g(\cdot)$ we obtain via (2) the following expression:

$$R(\delta,\theta) = 2Eg^2(\chi^2_{(p+2,\theta'\theta)}) + \theta'\theta[Eg^2(\chi^2_{(p+4,\theta'\theta)}) - 2Eg(\chi^2_{(p+2,\theta'\theta)}) + 1] \qquad (3)$$

where $\chi^2_{(j,\lambda)}$ denotes a chi-square random variable with j degrees of freedom and noncentrality parameter λ.

Lemma 1: Assume it is not true that $g(\cdot)=1$ a.e.. There is $\lambda_g \geq 0$ such that $0 < 2(Eg^2(\chi^2_{(p+2,\lambda_g)})-1) + \lambda_\delta[Eg^2(\chi^2_{(p+4,\lambda_g)}) - 2Eg(\chi^2_{(p+2),\lambda_g)}) + 1]$ if $p \geq 2$.

Proof: If p=2 and $D=I_2$, an estimator $\delta_{(n)}(X) = g(X'X+\chi^2_{(n)})X$, where $\chi^2_{(n)}$ is a random variable with chi-square (n) distribution independent of X (let $\chi^2_{(n)} \equiv 0$ if n=0), has in view of (3) risk
$R(\delta_{(n)},\theta) = 2Eg^2(\chi^2_{(4+n,\theta'\theta)})+ \theta'\theta[Eg^2(\chi^2_{(6+n,\theta'\theta)})-2Eg(\chi^2_{(4+n,\theta'\theta)}) +1]$.

For p=2 and $D=I_2$, Stein [5] has shown that the essentially unique estimator $\delta_0(X)=X$ is minimax admissible with constant risk $R(\delta_0,\theta)=2$ for all θ. If it is not true that $g(\cdot)=1$ a.e., then there exists a number $\lambda_{(g,n)} \geq 0$ such that $0 < R(\delta_{(n)},\theta)-R(\delta_0,\theta) = 2(Eg^2(\chi^2_{(4+n,\lambda_{(g,n)})})-1)$
$+\lambda_{(g,n)}[Eg^2(\chi^2_{(6+n,\lambda_{(g,n)})})-2Eg(\chi^2_{(4+n,\lambda_{(g,n)})}) +1]$, for all θ such that $\theta'\theta = \lambda_{(g,n)}$.

Let n=p-2 and $\lambda_g=\lambda_{(g,p-2)}$ so that there exists $\lambda_g \geq 0$ such that
$0 < 2(Eg^2(\chi^2_{(p+2,\lambda_g)})-1)+\lambda_g[Eg^2(\chi^2_{(p+4,\lambda_g)})-2Eg(\chi^2_{(p+2,\lambda_g)})+1]$.

q.e.d.

Lemma 2: If $\hat\theta(X)=(I_p-h(X'D^{-1}X)C)X$ is minimax and it is not true that $h(\cdot)=0$ a.e., then $E[g^2(\chi^2_{(p+2,\lambda_g)})] < 1$ where λ_g is given in Lemma 1 and $g(\cdot)=1-c_s h(\cdot)$, c_s being the smallest characteristic root of C, and $p > 2$.

Proof: Define $C_1=c_s^{-1}C$ and g as above. $\hat\theta$ is minimax if and only if $R(\hat\theta,\theta) \leq trD$ for all θ (because $\hat\theta_0(X)=X$ is minimax and $R(\hat\theta_0,\theta)=trD$, for all θ). Thus $\hat\theta$ minimax implies

$0 \geq R(\hat\theta,\theta)-trD = tr(C_1^2D)E[(1-g(\chi^2_{(p+2,\theta'\theta D^{-1}\theta)}))^2]$
$+ (\theta'C_1^2\theta)E[(1-g(\chi^2_{(p+4,\theta'D^{-1}\theta)}))^2]-2(\theta'C_1\theta)E[1-g(\chi^2_{(p+4,\theta'D^{-1}\theta)})]$
$-2tr(C_1D)E[1-g(\chi^2_{(p+2,\theta'D^{-1}\theta)})] + 2(\theta'C_1\theta)E[1-g(\chi^2_{(p+2,\theta'D^{-1}\theta)})]$ (by (2))
$\geq tr(C_1D)\{E[g^2(\chi^2_{(p+2,\theta'D^{-1}\theta)})] -1\} + (\theta'C_1\theta)\{E[g^2(\chi^2_{(p+4,\theta'D^{-1}\theta)})]$
$- 2E[g(\chi^2_{(p+2,\theta'D^{-1}\theta)})] + 1\}$

(because $tr(C_1^2D) \geq c_s^{(1)}tr(C_1D)$ and $\theta'C_1^2\theta \geq c_s^{(1)}(\theta'C_1\theta)$ where $c_s^{(1)} = 1$ is the smallest characteristic root of C_1.)

Choose θ_1 such that $\theta_1'D^{-1}\theta_1=\lambda_g$ as in Lemma 1 and $\theta_1'C_1\theta_1=\omega_s\lambda_g$ where ω_s is the smallest characteristic root of $D^{\frac{1}{2}}C_1D^{\frac{1}{2}}$. Then substituting above, we have $0 \geq R(\hat\theta,\theta_1)-trD \geq tr(C_1D)\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\}$
$+ \omega_s\lambda_g\{E[g^2(\chi^2_{(p+4,\lambda_g)})] - 2E[g(\chi^2_{(p+2,\lambda_g)})]+1\}$.

Assume $E[g^2(\chi^2_{(p+2,\lambda_g)})] \geq 1$. Then $0 \geq R(\hat{\theta},\theta_1)-\mathrm{tr}D \geq \omega_s[2\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\} + \lambda_g\{E[g(\chi^2_{(p+4,\lambda_g)})]-2E[g(\chi^2_{(p+2,\lambda_g)})]+1\}$ (because $\mathrm{tr}C_1D \geq p\omega_s > 2\omega_s$ and $\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\} \geq 0$).

But $\omega_s[2\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\} + \lambda_g\{E[g^2(\chi^2_{(p+4,\lambda_g)})]-2E[g(\chi^2_{(p+2,\lambda_g)})]+1\}] > 0$, a contradiction. Thus $E[g^2(\chi^2_{(p+2,\lambda_g)})] < 1$.

q.e.d.

Theorem 6: Let $X \sim N(\theta, D)$ and $\hat{\theta}(X)=(I_p-h(X'D^{-1}X)C)X$ where C,D are positive definite (symmetric pxp matrices and $h:[0,\infty) \to (-\infty,+\infty)$. Let ξ_L be the largest characteristic root of $D^{\frac{1}{2}}CD^{\frac{1}{2}}$. If $\mathrm{tr}CD \leq 2\xi_L$, then no estimator of the form $\hat{\theta}$ is minimax unless $h(\cdot)=0$ a.e..

Proof: Assume that it is not true that $h(\cdot)=0$ a.e.. Define C_1 as in proof of Lemma 2 and assume $\hat{\theta}$ is minimax; for all θ

$0 \geq R(\hat{\theta},\theta)-\mathrm{tr}D \geq \mathrm{tr}(C_1D)\{E[g^2(\chi^2_{(p+2,\theta'D^{-1}\theta)})]-1\}+(\theta'C_1\theta)\{E[g^2(\chi^2_{(p+4,\theta'D^{-1}\theta)})]$
$-2E[g(\chi^2_{(p+2,\theta'D^{-1}\theta)})]+1\}$ (as in proof of Lemma 2). Choose θ_1 such that $\theta_1'D^{-1}\theta_1=\lambda_g$ given in Lemma 1 and $\theta_1'C_1\theta_1=\omega_L\lambda_g$ where ω_L is the largest characteristic root of $D^{\frac{1}{2}}C_1D^{\frac{1}{2}}$. Assume $\mathrm{tr}CD \leq 2\xi_L$. Then substituting θ_1 for θ above, we have

$0 \geq R(\hat{\theta},\theta_1)-\mathrm{tr}D \geq \mathrm{tr}(C_1D)\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\}+\omega_L\lambda_g\{E[g^2(\chi^2_{(p+4,\lambda_g)})]$
$- 2E[g(\chi^2_{(p+2,\lambda_g)})]+1\} \geq \omega_L[2\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\}+\lambda_g\{E[g^2(\chi^2_{(p+4,\lambda_g)})]$
$-2E[g(\chi^2_{(p+2,\lambda_g)})]+1\}]$ (because Lemma 2 implies $\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\} < 0$ and because $\mathrm{tr}CD \leq 2\xi_L$ implies $\mathrm{tr}C_1D \leq 2\omega_L$.) But $\omega_L[2\{E[g^2(\chi^2_{(p+2,\lambda_g)})]-1\}$ $+\lambda_g\{E[g^2(\chi^2_{(p+4,\lambda_g)})]-2E[g(\chi^2_{(p+2,\lambda_g)})]+1\}] > 0$ by Lemma 1. This is a contradiction to the above and thus $\hat{\theta}$ minimax implies $\mathrm{tr}CD > 2\xi_L$.

q.e.d.

For the case $C=I_p$, this result has been shown by Brown [3] using a different method of proof.

IV. THE EXISTENCE OF PROPER BAYES MINIMAX ESTIMATORS

In Section II, examples were given of proper Bayes minimax estimators of the form $\hat{\theta}(X)=(I_p-Ch(X'D^{-1}X))X$ when $C=I_p$ and $\frac{\mathrm{tr}CD}{\xi_L} > \frac{p}{2}+2$. In this section it will be shown that when $C=I_p$ and $\frac{\mathrm{tr}CD}{\xi_L} \leq \frac{p}{2} + 2$, no proper Bayes minimax estimator of the form $\hat{\theta}$ exists.

First a bound for the bias of an estimator of the form $\hat{\theta}$ is given. An estimator of the form $\hat{\theta}(X)=(I_p-Ch(X'D^{-1}X))X$ has bias $E_\theta[\delta(X)]-\theta = -E[h(X'D^{-1}X)CX] = -E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})]C\theta$ (by Theorem 2).

Theorem 7: If $\hat{\theta}(X)=(I_p-h(X'D^{-1}X)C)X$ is a minimax estimator and $C=I_p$ and $\mathrm{tr}CD > 2\xi_L$, then

$$0 \leq E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})] \leq \frac{2\left(\frac{trCD}{\xi_L} -2\right)}{(\theta'D^{-1}\theta)} .$$

Proof: Because $\hat{\theta}$ is minimax, $0 \geq R(\hat{\theta},\theta)-trD = E[h^2(X'D^{-1}X)X'C^2X]$ $-2E[h(X'D^{-1}X)X'CX] + 2\theta'E[h(X'D^{-1}X)CX]$ (by (1)) $\geq (E[h(X'D^{-1}X)CX])'(E[h(X'D^{-1}X)CX])$ $-2E[h(X'D^{-1}X)X'CX] + 2\theta'E[h(X'D^{-1}X)CX]$ (by Jensen's inequality) $=$ $(\theta'C^2\theta)(E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})])^2-2trCD\ E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})]$ $-2(\theta'C\theta)E[h(\chi^2_{(p+4,\theta'D^{-1}\theta)})] + 2(\theta'C\theta)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})]$ (by Theorems 2 and4) For a given value of $\theta'D^{-1}\theta$, choose θ such that $\theta'C^2\theta=\theta'C\theta=\xi_L\theta'D^{-1}\theta$ where ξ_L is the largest characteristic root of $D^{\frac{1}{2}}CD^{\frac{1}{2}}$. The above inequality implies for a given value of $\theta'D^{-1}\theta$, $0 \geq (\theta'D^{-1}\theta)\xi_L(E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})])^2 -$ $2trCD\ E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})]-2\xi_L(\theta'D^{-1}\theta)E[h(\chi^2_{(p+4,\theta'D^{-1}\theta)})] +$ $2\xi_L(\theta'D^{-1}\theta)E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})]$.

Define $\psi(\lambda)= E[h(\chi^2_{(p+2,\lambda)})]$. The above inequality implies that for $\lambda \geq 0$, $\frac{d}{d\lambda}(\psi(\lambda)) \geq \frac{\psi(\lambda)}{4\lambda}\left(\psi(\lambda) - \frac{2(trCD-2\xi_L)}{\xi_L}\right)$

Replacing $2(p-2)$ by $\frac{2(trCD-2\xi_L)}{\xi_L}$ in the proof of Lemma I of Strawderman [8], p. 52, we have that for $p > 2$, $0 \leq \psi(\lambda) \leq \frac{2(trCD-2\xi_L)}{\xi_L}$.

Thus, $0 \leq E[h(\chi^2_{(p+2,\theta'D^{-1}\theta)})] \leq \frac{2(trCD-2\xi_L)}{\xi_L\theta'D^{-1}\theta}$. q.e.d.

An estimator of the form $\hat{\theta}(X)=(I_p-Ch(X'D^{-1}X))X$ has the form $g(X'D^{-1}X)X$ when $C=I_p$ and $g(\cdot)=1-h(\cdot)$.

Theorem 8: If $\frac{trD}{\xi_L} \leq \frac{p}{2}+2$, no estimator of the form $g(X'D^{-1}X)X$ is proper Bayes minimax where $X \sim N(\theta,D)$ and $g:[0,\infty) \to (-\infty,+\infty)$.

Proof: Assume $\hat{\theta}(X)=g(X'D^{-1}X)X$ is a minimax proper Bayes estimator for θ under the loss function $(\hat{\theta}-\theta)'(\hat{\theta}'\theta)$ with respect to the prior density $dH(\theta)$. Let $Y=D^{-\frac{1}{2}}X \sim N(\eta,I_p)$ where $\eta=D^{-\frac{1}{2}}\theta$. Then $\hat{\eta}(Y)=D^{-\frac{1}{2}}\hat{\theta}(D^{\frac{1}{2}}Y)$ is the minimax proper Bayes estimator for η under the loss function $(\hat{\eta}-\eta)'D(\hat{\eta}-\eta)$ with respect to the prior density $dG(\eta)$ where G is the distribution generated by H under the transformation $\eta=D^{-\frac{1}{2}}\theta$.

$\hat{\theta}(X)$is of the form $g(X'D^{-1}X)X$ if and only if $\hat{\eta}$ is of the form $g(Y'Y)Y$, and it has been shown by Strawderman and Cohen [6] that this is true if and only if G is invariant under orthogonal transformations.

Brown [4] has noted, p. 858, that

$$\hat{\eta}_i(Y) - Y_i = \frac{\partial}{\partial Y_i}\left[\log\left(\int_{\mathcal{R}^p} \frac{e^{-\frac{(Y-\eta)'(Y-\eta)}{2}}}{(2\pi)^{p/2}}\, d G(\eta)\right)\right]$$

Defining $h(\cdot)=1-g(\cdot)$, we have $\hat{\eta}_i(Y)-Y_i = -h(Y'Y)Y_i$. By Strawderman [8],

$$\frac{\partial}{\partial \eta_i}\left(\int_{\mathcal{R}^p}\left[\log\left(\int_{\mathcal{R}^p} \frac{e^{-\frac{1}{2}(Y-\eta)'(Y-\eta)}}{(2\pi)^{p/2}}\, dG(\eta)\right)\right] \frac{e^{-\frac{1}{2}(Y-\eta)'(Y-\eta)}}{(2\pi)^{p/2}}\, dy\right)$$

$= E_\eta[\hat{\eta}_i(Y) - Y_i] = -E[h(\chi^2_{(p+2,\eta'\eta)})]\eta_i$ (by Theorem 1)

$\geq \dfrac{-2\left(\frac{trCD}{\xi_L} - 2\right)\eta_i}{\eta'\eta}$ for $\eta_i > 0$ (by Theorem 7 since $\eta\ \eta = \theta' D^{-1}\theta$).

Using this inequality and the fact that G is orthogonally invariant, the proof of Theorem 2 of Strawderman [8], p. 54, implies that for $C=I_p$ $2(\frac{trCD}{\xi_L} - 2) - (p-1) > 1$, i.e., $\frac{trCD}{\xi_L} > \frac{p}{2} + 2$. Thus if $\hat{\theta}(X)$ is a spherically symmetric minimax Bayes estimator $\frac{trCD}{\xi_L} > \frac{p}{2} + 2$.

q.e.d.

REFERENCES

[1] Baranchik, A.J., "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Stanford University Technical Report No. 51.

[2] Bock, M.E., "Minimax Estimators of the Mean of a Multivariate Normal Distribution," Annals of Statistics (1973), to appear.

[3] Brown, L., "Estimation with Incompletely Specified Loss Function (the Case of Several Location Parameters," Journal of the American Statistical Association (1973), to appear.

[4] Brown, L., "Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems," Annals of Mathematical Statistics, 42(1971), 855-903.

[5] Stein, C., "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley (1950), 197-206.

[6] Strawderman, W.E. and Cohen, A., "Admissibility of Estimators of the Mean Vector of a Multivariate Normal Distribution with Quadratic Loss," Annals of Mathematical Statistics, 42(1971), 270-296.

[7] Strawderman, W.E., "Proper Bayes Minimax Estimators of the Multivariate Normal Mean," Annals of Mathematical Statistics, 42(1971), 385-388.

[8] Strawderman, W.E., "On the Existence of Proper Bayes Minimax Estimators of the Mean of a Multivariate Normal Distribution," Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, (1972), 51-55.

THE CHARACTERISTIC FUNCTION AND THE CORE OF A MULTISTAGE GAME OF COALITIONS

A. HAURIE**

ECOLE DES HAUTES ETUDES COMMERCIALES

MONTREAL

1. Introduction:

In this paper some properties of the Core and of the Characteristic Function of a multistage m-player game without side payments are explored. The Core of a m-player game has been extensively studied for games in characteristic function form; at this level of abstraction the only givens of the game are the payoffs which each of the possible coalitions of players can assureffor themselves respectively (see for example [1] - [3]). Most of the results obtained deal with the non emptiness of the Core; an important assumption in these theories is the convexity and the closedness of the characteristic function for a given coalition.

For a game in normal form, which is represented by the strategies available to each player and the payoffs corresponding to any possible choice of strategies, there has been recently some work to characterize the Core or the Characteristic Function ([4] - [6]).

In this paper one deals with games in extensive form with perfect information, which can be represented by a multistage control system with strategies described by feedback control. For this class of games, using an approach inspired from the control of perturbed systems ([7] - [10]), one defines the Characteristic Function and the Core. Then it will be shown that the concept of Core is not satisfactory since the multistage-feedback structure of the game requires a more restrictive definition of optimality to provide collective rationality. The concept of C-optimal control sequence is proposed and a dynamic programming algorithm permits the characterization of such sequences when the Core of the game has been defined in each state at each stage. This will complete the results [11]

When the dynamical system is described by a linear difference equation and the cost functionals are convex with some particular form it is shown that the construction of the characteristic function can be obtained in a manner similar to the method used by D. Bertsekas [8] for Minimax Control problems; this dynamic programming algorithm is used to show that the characteristic function has the desirable properties of closedness and biconvexity. This means that not only the set of payoffs which a coalition can assure for itself at given state and stage is closed and convex but also that there exist similar properties when one considers the characteristic function at different states for a given stage.

Abstract: The Characteristic Function and the Core of a game in extensive form, described by a multistage control system, are studied. It is shown that the concept of Core must be strengthened to provide collective rationality in a multistage game with feedback strategies. Dynamic programming algorithms are proposed to compute the optimal control sequences and to obtain important convexity and closedness properties of the characteristic function of a game described by a linear-convex system.

---

* This work was supported by the Canada Council (S72-0513) and by the Ministère de l'Education du Québec (DGES).

2. The Characteristic Function, the Core, and C-optimal Control sequences for a Multistage Game of Coalitions.

Consider the multistage dynamic system

$$x(k+1) = f^k[x(k),\ u_M(k)],\ k = 0,\dots,\ K-1 \tag{2.1}$$

where $x(k) \in E^n$, $k = 0, 1,\dots, K$ is the state vector, $M \triangleq \{1, 2,\dots, m\}$ the set of players,

$$u_j(k) \in U_j^k \subseteq E^{r_j} \quad \text{and} \quad u_M(k) \in \prod_{j \in M} U_j^k,\ k = 0, 1,\dots, K-1$$

are the control vectors of each player $j \in M$ and of the set of all players M respectively.

A coalition S is a subset of M. For a given coalition S denote:

$$u_S(k) \triangleq \{u_j(k)\}_{j \in S} \in U_S^k \triangleq \prod_{j \in S} U_j^k \tag{2.2}$$

$$u_{M-S}(k) \triangleq \{u_j(k)\}_{j \in M-S} \in U_{M-S}^k \tag{2.3}$$

If a coalition $S \neq M$ forms itself it considers strategies of the form:

$$\begin{aligned} \mu_S^k : E^n &\to U_S^k, \quad k = 0, 1,\dots, K-1 \\ x(k) &\to \mu_S^k[x(k)] \end{aligned} \tag{2.4}$$

and a vector valued cost function defined by:

$$\psi_S[\cdot] \triangleq \{\psi_{j/S}[\cdot]\}_{j \in S} \tag{2.5}$$

with,

$$\psi_{j/S}[i,\ x(i),\ \mu_S^i,\dots,\ \mu_S^{K-1}] \triangleq$$

$$\sup_{\substack{u_{M-S}(k) \in U_{M-S}^k \\ k=i,\ i+1,\dots,\ K-1}} \left\{ \sum_{k=i}^{K-1} \phi_j^k[x(k),\ \mu_S^k[x(k)],\ u_{M-S}(k)] + \Theta_j[x(K)] \right\} \tag{2.6}$$

where $\phi_j^k : E^n \times U_M^k \to (-\infty, \infty]$ and $\Theta_j : E^n \to (-\infty, \infty]$ are given functionals.

Thus, by Eq(2.6), the cost function used for the evaluation of the strategies of a coalition S depends on the initial stage i, the initial state $x(i)$ and the strategies used at each subsequent stage. Furthermore when the player j is in the coalition S then he knows only the strategies of this coalition and, adopting a pessimistic approach, he considers the controls of the players in the complementary coalition M-S as unpredictable, but constrained, perturbations.

When the coalition S is equal to M , one has full cooperation and for each player the following cost functional is defined:

$$J_j[i, x(i), u_M(i),\ldots, u_M(K-1)] \triangleq \sum_{k=i}^{K-1} \phi_j^k[x(k), u_M(k)] + \Theta_j[x(K)] \qquad (2.7)$$

Definition 2.1: The Characteristic Function (at stage i) of the game of coalitions given by Eqs (2.1) - (2.7) is the mapping v^i defined by:

$$v^i : (S, x(i)) \rightarrow v^i[S, x(i)] \subset R^S \triangleq \prod_{j \in S} (-\infty, \infty]$$

with,

$$\omega_S \triangleq \{\omega_j\}_{j \in S} \in v^i[S, x(i)] \Leftrightarrow \exists\, \mu_S^i, \ldots, \mu_S^{K-1}$$

$$\text{s.t.} \quad \forall_j \in S \quad \psi_{j/S}[i, x(i), \mu_S^i, \ldots, \mu_S^{K-1}] \leq \omega_j \qquad (2.8)$$

Thus the characteristic function gives for each stage i , each state x(i) and each coalition S the set of all cost vectors ω_S which are dominated by admissible values of the cost function (2.5).

Now, for the sake of generality, consider the Coalition Constraints defined by the correspondences

$$\Gamma^k : x(k) \rightarrow \Gamma^k[x(k)] \subset \mathcal{P}(M) \quad k = 0, 1, \ldots, K \qquad (2.9)$$

where $\mathcal{P}(M)$ is the class of all subsets of M . A coalition $S \in \Gamma^k[x(k)]$ will be called an admissible Coalition at x(k) .

Definition 2.2: The Core at x(i) of the game given by Eqs (2.1) - (2.9) is the set $\Omega^i[x(i)]$ of all vectors $\omega_M \in R^M$ such that

(i) There exists a Pareto-optimal sequence of controls

$$u_M(i), \ldots, u_M(K-1)$$

such that

$$\forall j \in M \quad J_j[i, x(i), u_M(i), \ldots, u_M(K-1)] = \omega_j \qquad (2.10)$$

(ii) $\forall S \in \Gamma^i[x(i)]$ the projection ω_S of ω_M in R^S is not interior to $v^i[S, x(i)]$.

If a cost vector ω_M is in the core $\Omega^i[x(i)]$ then it is an undominated outcome when full cooperation is realized and furthermore, at this point x(i) , no admissible coalition can assure lower cost levels for all its members independently of the actions of the complementary coalition. In other words, no coalition can rise an objection against the control sequence at this point. Generally games of coalitions are defined directly in a characteristic function form (Refs. [1] - [4]); thus the Core is a usefull solution concept having some equilibrium properties. However, an important point does not seem to have been noticed yet, when dealing with games defined in extensive form (Refs.[13 [14): Even if a cost vector ω_M is in the Core $\Omega^i[x(i)]$, this does not provide the assurance that at a later stage k, at a point $x(k)$ of the trajectory generated by the Pareto-optimal controls, verifying (2.10), no admissible coalition will be able to obtain, by its own action, a "pessimistic" cost vector dominating the cost vector generated by the controls $u_M(k),\dots, u_M(K-1)$. Hence a more restrictive definition of optimality is needed.

Definition 2.3: A control sequence $\hat{u}_M(k)$, $k=i, i+1,\dots, K-1$ is C-optimal at $\hat{x}(i)$ if it generates a trajectory $\hat{x}(k)$, $k=i, i+1,\dots, K$, solution of Eq (2.1) and a sequence of costs:

$$\omega_j^k \triangleq J_j[k, \hat{x}(k), \hat{u}_M(k),\dots, \hat{u}_M(K-1)] \qquad (2.11)$$

$$j \in M, \quad k = i, i+1,\dots, K-1$$

such that

$$\omega_M^k \triangleq \{\omega_j^k\}_{j \in M} \in \Omega^k[\hat{x}(k)], \quad k = i, i+1,\dots, K-1 \qquad (2.12)$$

■

Of course, a C-optimal control sequence is also Pareto Optimal; furthermore, since the cost vector ω_M^k is contained at each state $\hat{x}(k)$ in the Core $\Omega^k[\hat{x}(k)]$ there is a collective rationality (Refs [11]) in using the control sequence

$$\hat{u}_M(i), \hat{u}_M(i+1),\dots, \hat{u}_M(K-1)$$

The following notations will be useful:

$$\varphi_S^k[\cdot] = \{\varphi_j^k[\cdot]\}_{j \in S}$$

$$J_S[\cdot] = \{J_j[\cdot]\}_{j \in S}$$

A Dynamic Programming approach can be used to characterize C-optimal control sequences when the Core of the game is known at each stage k and for each stage $x(k)$.

Theorem 2.1: Define in E^{n+m} the sets $Y(k)$ and the mappings $p_M^k : Y(k) \to U_M^k$ by the following rules:

(i) $Y(K-1) \triangleq \{(x(K-1), \omega_M(K-1)) : \exists\, \bar{u}_M(K-1) \triangleq$
$p_M^{K-1}[x(K-1), \omega_M(K-1)] \in U_M^{K-1}$ s.t.
$\omega_M(K-1) = J_M[K-1, x(K-1), \bar{u}_M(K-1)] \in \Omega^{K-1}[x(K-1)]\}$ (2.13)

(ii) $Y(k) \triangleq \{(x(k), \omega_M(k)) : \exists\, \bar{u}_M(k) \triangleq p_M^k[x(k), \omega_M(k)] \in U_M^k$

such that $\omega_M(k) \in \Omega^k[x(k)]$ and

$$(f^k[x(k), \bar{u}_M(k)], \omega_M(k) - \phi_M^k[x(k), \bar{u}_M(k)]) \in Y(k+1)\} \qquad (2.14)$$

for $k = K-2, \dots, 0$;

then there exists a C-optimal sequence at $\hat{x}(i)$ generating the cost $\hat{\omega}_M(i) \in E^m$ iff $(\hat{x}(i), \hat{\omega}_M(i)) \in Y(i)$. Furthermore the strategies $p_M^k : Y(k) \rightarrow U_M^k$ generate such a C-optimal control sequence at $\hat{x}(i)$.

Proof: If $(\hat{x}(i), \hat{\omega}(i)) \in Y(i)$ then by (2.13) and (2.14) the strategies p_M^k, $k = i, \dots, K-1$ will generate a trajectory and control sequence verifying (2.11) and (2.12) therefore, the control sequence will be C-optimal at $\hat{x}(i)$.

If $(\hat{x}(i), \hat{\omega}(i)) \notin Y(i)$ then either:

$$\hat{\omega}(i) \notin \Omega^i[(\hat{x}(i)] \qquad (2.15)$$

or

$$(f^i[\hat{x}(i), u_M(i)], \hat{\omega}(i) - \phi_M^i[\hat{x}(i), u_M(i)])$$

$$\notin \; Y(i+1) \quad , \quad \forall u_M(i) \in U_M(i) \qquad (2.16)$$

If (2.15) is verified then there does not exist a C-optimal control sequence at $\hat{x}(i)$ generating the cost vector $\hat{\omega}_M(i)$.

Assume that, at stage $(i+1)$, $(x(i+1), \omega_M(i+1)) \notin Y(i+1)$ implies that there does not exist a C-optimal control sequence a $x(i+1)$ generating the cost vector $\omega_M(i+1)$; then (2.15) and (2.16) show that the same property will be true at stage i .

Since the property holds at stage $K-1$ it is true at any stage. ■

To define collective rationnality for a multistage game a concept more restrictive than the Core is needed.

Definition 2.4: The Super-Core of the game at $x(k)$ is the subset $\overset{+}{\Omega}{}^k[x(k)]$ of $\Omega^k[x(k)]$ defined by: $\omega_M \in \overset{+}{\Omega}{}^k[x(k)] \Leftrightarrow \exists$ a C-optimal control sequence $\hat{u}_M(k), \dots, \hat{u}_M(K-1)$ such that

$$\omega_M = J_M[k, x(k), \hat{u}_M(k), \dots, \hat{u}_M(K-1)]$$

Theorem 2.2: The Super-Core at $x(k)$ is given by the following rules

(i) $\overset{+}{\Omega}{}^{K-1}[x(K-1)] = \Omega^{K-1}[x(K-1)]$ (2.17)

(ii) $\overset{+}{\Omega}{}^k[x(k)] \triangleq \{\omega_M^k \in \Omega^k[x(k)] : \exists\, \bar{u}_M(k)$ and $\omega_M^{k+1} \in$

$$\overset{+}{\Omega}{}^{k+1}[f^k[x(k), \bar{u}_M(k)]] \text{ s.t. } \omega_M^k = \phi_M^k[x(k), \bar{u}_M(k)] + \omega_M^{k+1}\} \qquad (2.18)$$

$k = K - 2, \ldots, 0$.

Proof: Same as in theorem 2.1. ■

3. Properties of the characteristic function of a linear-convex game.

In this section a linear multistage system is considered:

$$x(k+1) = A^k x(k) + \sum_{j \in M} B_j^k u_j(k) \tag{3.1}$$

with the following particular cost functionals:

$$\psi_{j/S}[i, x(i), \mu_S^i, \ldots, \mu_S^{K-1}] \triangleq$$

$$\sup_{\substack{u_{M-S}(k) \in U_{M-S}^k \\ k=i,\, i+1, \ldots, K-1}} \left\{ \sum_{k=i}^{K-1} \left(\varphi_j^k[x(k)] + \gamma_j^k[\mu_{j/S}^k(x(k))] \right) + \varphi_j^K[x(K)] \right\} \tag{3.2}$$

where A^k, B_j^k are given matrices and the functionals

$\varphi_j^k : E^n \to (-\infty, \infty]$, $k = 0, 1, \ldots, K$, $\gamma_j^k : U_j^k \to (-\infty, \infty]$, $k = 0, 1, \ldots, K-1$

are given *closed proper convex functionals*.

$\mu_{j/S}^k[x(k)] \in U_j^k$ is the projection on U_j^k of the vector $\mu_S^k[x(k)]$.

Of course, the control sets U_j^k are assumed to be convex. This game is an adaptation to a m-player game setting of the uncertain system extensively studied in Refs [8] - [ ], by D. Bertsekas. Now it is possible to give a constructive method for the determination of the characteristic function of the game. The following notations will be usefull:

$$\omega_S \triangleq \{\omega_j\}_{j \in S} \quad , \quad \varphi_S^k(.) \triangleq \{\varphi_j^k(.)\}_{j \in S}$$

$$\omega_S \le \omega_S' \quad \Leftrightarrow \quad \omega_j \le \omega_j' \quad \forall j \in S$$

$$B_S^k u_S(k) \triangleq \sum_{j \in S} B_j^k u_j(k)$$

Theorem 3.1: The characteristic function of the linear convex game given by Eqs (3.1) and (3.2) can be constructed by the following algorithm:

1. $\forall x(K)$, $v^K[S, x(K)] \triangleq \{\omega_S \in R^S : \varphi_S^K[x(K)] \le \omega_S\}$ (3.3)

2. $\mathcal{E}^{k+1}[S, x] \triangleq \bigcap_{u_{M-S}(k) \in U_{M-S}^k} v^{k+1}[S, x + B_{M-S}^k u_{M-S}(k)]$ (3.4)

3. $\mathcal{H}^k[S, x(k)] \triangleq \bigcup_{u_S(k) \in U_S^k} \{\gamma_S^k[u_S(k)]\} + \mathcal{E}^{k+1}[S, A^k x(k) + B_S^k u_S(k)]$ (3.5)

4. $v^k[S, x(k)] = \{\varphi_S^k[x(k)]\} + \mathcal{H}^k[S, x(k)]$ (3.6)

Proof: First the following proposition will be established:

$$\omega_S \in v^k[S,\ x(k)] \Leftrightarrow \exists\ \bar{u}_S(k) \in U_S^k \quad \text{s.t.}$$

$$\forall u_{M-S}(k) \in U^k_{M-S}, \omega_S - \varphi^k_S[x(k)] - \gamma^k_S[\bar{u}_S(k)]$$

$$\in v^{k+1}[S,\ A^k\ x(k) + B^k_S\ \bar{u}_S(k) + B^k_{M-S}\ u_{M-S}(k)] \tag{3.7}$$

Necessary condition: Consider $\omega_S \in v^k[S,\ x(k)]$, thus there exist $\mu^k_S, \dots, \mu^{K-1}_S$ such that

$$\psi_S[k,\ x(k)\ ,\ \mu^k_S, \dots, \mu^{K-1}_S] \le \omega_S \tag{3.8}$$

Defining $\bar{u}_S(k) \triangleq \mu^k_S[x(k)]$, the R.H.S. of (3.7) is trivialy verified.

Sufficient condition: Consider $\bar{u}_S(k)$ such that the R.H.S. of (3.7) is verified, then for any $u_{M-S}(k)$ there exists a sequence of strategies $\mu^{k+1}_S, \dots, \mu^{K-1}_S$ such that

$$\psi_S[k+1,\ \bar{x}(k+1),\ \mu^{k+1}_S, \dots, \mu^{K-1}_S] \le \omega'_S \tag{3.9}$$

with
$$\omega'_S \triangleq \omega_S - \varphi^k_S[x(k)] - \gamma^k_S[\bar{u}_S(k)] \tag{3.10}$$

$$\bar{x}(k+1) \triangleq A^k\ x(k) + B^k_S\ \bar{u}_S(k) + B^k_{M-S}\ u_{M-S}(k) \tag{3.11}$$

Thus, defining $\mu^k_S[x(k)] \triangleq \bar{u}_S(k)$ one has:

$$\forall\ j \in S$$

$$\psi_{j/S}[k,\ x(k),\ \mu^k_S, \dots, \mu^{K-1}_S] = \varphi^k_j[x(k)] + \gamma^k_S[\bar{u}_S(k)]$$

$$+ \sup_{u_{M-S}(k) \in U^k_{M-S}} \psi_{j/S}[k+1,\ \bar{x}(k+1)\ ,\ \mu^{k+1}_S, \dots, \mu^{K-1}_S] \tag{3.12}$$

and then by Eqs (3.9) - (3.12):

$$\psi_S[k,\ x(k)\ ,\ \mu^k_S, \dots, \mu^{K-1}_S] \le \omega_S \tag{3.13}$$

that is, by Definition 2.1: $\omega_S \in v^k[S,\ x(k)]$. This establishes the proposition (3.7).

Now it remains to show that the three last steps of the algorithm are equivalent to the R.H.S. of (3.7). By Eq (3.6), one has:

$$\omega_S \in v^k[S,\ x(k)] \Leftrightarrow \omega_S - \varphi^k_S[x(k)] \in \mathcal{H}^k[S,\ x(k)] \tag{3.14}$$

and by Eqs (3.5) and (3.14):

$$\omega_S \in v^k[S,\ x(k)] \Leftrightarrow \exists\ u_S(k) \in U^k_S \quad \text{s.t.}$$

$$\omega_S - \varphi^k_S[x(k)] - \gamma^k_S[\bar{u}_S(k)] \in \mathcal{E}^{k+1}[S,\ A^k\ x(k) + B^k_S\ u_S(k)] \tag{3.15}$$

and finally by Eqs (3.4) and (3.15):

$$\omega_S \in v^k[S,\, x(k)] \Leftrightarrow \exists\, u_S(k) \in U_S^k \quad \text{s.t.} \quad \forall\, u_{M-S}(k) \in U_{M-S}^k$$

$$\omega_S - \varphi_S^k[x(k)] - \gamma_S^k[\bar{u}_S(k)] \in v^{k+1}[S,\, A^k\, x(k) + B_S^k\, u_S(k) + B_{M-S}^k\, u_{M-S}(k)]$$

that is the R.H.S. of (3.7). ■

Theorem 3.2: The characteristic function of the linear convex game given by Eqs (3.1) and (3.2) is such that for any stage k and any coalition S the following property holds:

$$\forall\, x\,,\ x' \;\; E^n\,,\ \forall \alpha \in [0,\, 1]$$

$$\alpha\, v^k(S,\, x) + (1 - \alpha)\, v^k(S,\, x') \subset v^k(S, \alpha\, x + (1 - \alpha)\, x') \tag{3.16}$$

Proof: At stage K, the convexity of the function $\varphi_S^K[\circ]$ implies that (3.16) holds.

Assume now that (3.16) is true at stage $k+1$ then, using (3.4) one has:

$$\forall\, x\,,\ x' \;\; E^n\,,\ \forall \alpha \in [0,\, 1]$$

$$\alpha\, \mathcal{E}^{k+1}[S,\, x] + (1 - \alpha)\, \mathcal{E}^{k+1}[S,\, x'] \subset \mathcal{E}^{k+1}[S, \alpha\, x + (1 - \alpha)\, x'] \tag{3.17}$$

Now, for some x and x' consider two vectors

$$\omega_S \in \mathcal{H}^k[S,\, x]\,,\ \omega_S' \in \mathcal{H}^k[S,\, x']$$

By construction (Eq. (3.5)) there exist $u_S(k)$ and $u_S'(k)$ respectively such that:

$$\omega_S = \gamma_S(u_S(k)) + \lambda_S[x,\, u_S(k)]$$

$$\omega_S' = \gamma_S(u_S'(k)) + \lambda_S[x',\, u_S'(k)]$$

where $\quad \lambda_S[u,\, u_S(k)] \in \mathcal{E}^{k+1}[S,\, A^k x + B_S^k\, u_S(k)]$

and $\quad \lambda_S[u',\, u_S'(k)] \in \mathcal{E}^{k+1}[S,\, A^k x' + B_S^k\, u_S'(k)]$

Since $\gamma_S(\circ)$ is convex, the following holds: $\forall \alpha \in [0,\, 1]$

$$\begin{aligned} \alpha\, \omega_S + (1 - \alpha)\, \omega_S' \geq\ & \gamma_S[\alpha\, u_S(k) + (1 - \alpha)\, u_S'(k)] \\ & + \alpha\, \lambda_S[x,\, u_S(k)] + (1 - \alpha)\, \lambda_S[x',\, u_S'(k)] \end{aligned} \tag{3.19}$$

that is, by (3.17), (3.18), (3.19)

$$\alpha\, \omega_S + (1 - \alpha)\, \omega_S' \in \mathcal{H}^k[S, \alpha\, x + (1 - \alpha)\, x'] \tag{3.20}$$

Finally, (3.20) and Eq. (3.6) where $\varphi_S^k[\cdot]$ is convex imply

$$\forall\, x,\ x'\,,\ \forall \alpha \in [0,\, 1]$$

$$\alpha\, v^k[S,\, x] + (1 - \alpha)\, v^k[S,\, x'] \subset v^k[S, \alpha\, x + (1 - \alpha)\, x']$$

which completes the proof. ■

Theorem 3.3: The characteristic function of the linear convex game is such that: $\forall x \in E^n$, $\forall S \subset M$, $\forall k \in \{0, 1, \dots, K\}$, $v^k(S, x)$ is convex.

Proof: Let ω_S and ω'_S be two elements of $v^k(S, x)$, then by (3.6) and (3.5) there exist $u_S(k)$ and $u'_S(k)$ in U^k_S such that:

$$\omega_S = \varphi^k_S[x] + \lambda_S + \gamma^k_S[u_S(k)]$$

$$\omega'_S = \varphi^k_S(x) + \lambda'_S + \gamma^k_S[u'_S(k)]$$

with:

$$\lambda_S \in \mathcal{E}^{k+1}[S,\ A^k\ x + B^k_S\ u_S(k)]$$

$$\lambda'_S \in \mathcal{E}^{k+1}[S,\ A^k\ x + B^k_S\ u'_S(k)]$$

Thus, by (3.17) one has:

$$\forall\ \alpha\ \in [0,\ 1]\ \ \alpha\lambda_S + (1-\alpha)\ \lambda'_S \in \mathcal{E}^{k+1}[S,\ A^k\ x + B^k_S(\alpha\, u_S(k) + (1-\alpha)\ u'_S(k))]$$

and therefore by convexity of γ^k_S

$$\omega''_S \triangleq \alpha\,\omega_S + (1-\alpha)\ \omega'_S \geq \varphi^k_S(x) + \alpha\lambda_S + (1-\alpha)\ \lambda'_S + \gamma^k_S[\alpha\, u_S(k) + (1-\alpha)\ u'_S(k)]$$

i.e. ω''_S is an element of $v^k(S, x)$. ■ (Also take $x = x'$ in Theorem 3.2)

Theorem 3.4: Assume that the control sets U^k_j, $j \in M$, k $\{0,1,\dots,K-1\}$ are compact then the characteristic function of the linear convex game has the following properties: $\forall S \subset M$, $\forall k \in \{0, 1, \dots, K\}$

(i) $v^k[S, x]$ is closed $\forall x \in E^n$

(ii) If $(x_\sigma)_{\sigma \in N} \to \bar{x}$ and $(\omega_{S,\sigma})_{\sigma \in N} \to \bar{\omega}_S$ are two converging sequences in E^n and R^S respectively such that: $\forall \sigma \in N$ $\omega_{S,\sigma} \in v^k[S, x_\sigma]$ then $\bar{\omega} \in v^k[S, \bar{x}]$.

Proof: At stage K property (i) is obvious. By Eq. (3.3) and φ^K_j being closed proper convex one has:

$$\forall\ j \in S \quad \varphi^K_j(\bar{x}) \leq \lim_{\sigma \to \infty} \varphi^K_j(x_\sigma) \leq \lim\ \omega_{S,\sigma} \triangleq \bar{\omega}_S$$

Thus property (ii) holds at stage K. Assume that both properties hold at stage $k+1$ then, by Eqs (3.5) - (3.6):

$$\omega_{S,\sigma} \in v^k[S, x_\sigma] \Leftrightarrow \exists\, u_{S,\sigma}(k) \in U^k_S\ ,$$

$$\exists\, \lambda_{S,\sigma} \in \mathcal{E}^{k+1}[S,\ A^k x_\sigma + B^k_S\ u_{S,\sigma}(k)]\ \text{s.t.}$$

$$\omega_{S,\sigma} = \varphi^k_S[x_\sigma] + \gamma^k_S[u_{S,\sigma}(k)] + \lambda_{S,\sigma} \tag{3.21}$$

Since U^k_S is compact one may consider that the sequence $(u_{S,\sigma}(k))_{\sigma \in N}$ converges to a limit $\bar{u}_S(k)$. The functionals φ^k_S and γ^k_S being closed proper convex one has from Eq. (3.21):

$$\bar{\omega}_S \geq \varphi^k_S[\bar{x}] + \gamma^k_S[\bar{u}_S(k)] + \bar{\lambda}_S \tag{3.22}$$

where $\bar{\lambda}_S \triangleq \lim_{\sigma \to \infty} \lambda_{S,\sigma}$ is well defined by:

$$\bar{\lambda}_S = \bar{\omega}_S - \lim_{\sigma \to \infty} \{\varphi_S^k[x_\sigma] + \gamma_S^k[u_{S,\sigma}(k)]\} \tag{3.23}$$

Now it is readily verified that, if property (ii) holds at stage $k+1$, one has

$$\bar{\lambda}_S \in \mathcal{E}^{k+1}[S, A^k \bar{x} + B_S^k \bar{u}_S(k)] \tag{3.24}$$

and therefore, by Eqs (3.21) - (3.24), property (ii) holds at stage k. Property (i) is then a direct consequence of property (ii). ■

References

[1] H. SCARF, "The Core of a n-person game", Econometrica, n° 35, pp 50-69, 1967.

[2] R. AUMANN, "A Survey of Cooperative Games without Side Payments in Essays in Mathematical Economics", M. Shubik edit., Princeton, 1969.

[3] W.F. LUCAS, "Some recent developments in n-person game theory", SIAM Review, vol. 13, pp 491-523, 1971.

[4] H. SCARF, "On the existence of a cooperative solution for a general class of n-person games", Journal of Economic Theory, n° 3, pp 169-181, 1971.

[5] A. HAURIE, "On Pareto optimal decisions for a coalition of a subset of Players", IEEE Transactions on Automatic Control, vol. AC18, n° 2, April 1973, pp 144-149.

[6] J.L. GOFFIN and A. HAURIE, "Necessary condition and sufficient conditions for Pareto Optimality in a multicriterion perturbed system", To appear; proceedings of the 5th IFIP Conference on Optimization techniques, Rome 1973, Springer-Verlag.

[7] H.S. WITSENHAUSEN, "Minimax control of uncertain systems", Rep. ESL-R-269, MIT, May 1966.

[8] D.P. BERTSEKAS, "Control of uncertain systems with a set-membership description of the uncertainty", Rep. ESL-R-447, MIT, June 1971.

[9] D.P. BERTSEKAS and I.B. RHODES, "On the minimax reachability of Target Sets and Target Tubes" Automatica, Vol.7, 1971, pp. 233-247.

[10] D.P. BERTSEKAS and I.B. RHODES, "On the minimax feedback control of uncertain systems", IEEE Conference on decision and control, Proceeding paper, n° F3-3, 1971.

[11] A. HAURIE and M. DELFOUR, "Individual and Collective Rationality in a Dynamic Pareto Equilibrium", to appear in JOTA, March 1974.

[12] Von NEUMANN, J. and MORGENSTERN, O. "Theory of Games and Economic Behavior", Princeton Univ. Press, 1944.

[13] OWEN, G. "Game Theory", Saunders, 1968.

THE FEEDFORWARD AND FEEDBACK CONTROL OF A GENERAL SERVOMECHANISM PROBLEM

Part I

EDWARD J. DAVISON*
Department of Electrical Engineering
University of Toronto, Canada

ABSTRACT

Necessary and sufficient conditions are found for a linear time-invariant feedback or feedforward controller to exist for a general linear multivariable system subject to a general class of unmeasurable or/and measurable arbitrary disturbances, so that the outputs asymptotically track certain preassigned outputs and so that the controlled system is controllable (or stabilizable). Explicit feedforward and feedback controllers which will accomplish this are obtained. Some numerical examples are included to illustrate the results.

I. INTRODUCTION

Consider the following linear time-invariant system:

$$\begin{aligned} \dot{\underline{x}} &= \underline{A}\underline{x} + \underline{B}\underline{u} + \sum_{i=1}^{\Gamma} (\underline{E}_i, \underline{0}) \begin{bmatrix} \underline{\omega}_i \\ \\ \underline{y}_i \end{bmatrix} \\ \underline{e} &= \underline{C}\underline{x} + \underline{D}\underline{u} + \sum_{i=1}^{\Gamma} (\underline{F}_i, -\underline{G}_i) \begin{bmatrix} \underline{\omega}_i \\ \\ \underline{y}_i \end{bmatrix} \end{aligned} \tag{1}$$

where $\underline{x}$ is the n-state vector, $\underline{u}$ the m-control vector, $\underline{e}$ the r-error vector, $\underline{y} \triangleq \underline{C}\underline{x} + \underline{D}\underline{u} + \sum_{i=1}^{\Gamma} \underline{F}_i\underline{\omega}_i$ the r-output vector and $\underline{y}_{ref} \triangleq \sum_{i=1}^{\Gamma} \underline{G}_i\underline{y}_i$ is the desired specified asymptotic (as $t \to \infty$) r-output vector. $\underline{\omega}_i$ are n_i-disturbance vectors, which may or may not be measurable, where rank $\begin{bmatrix} \underline{E}_i \\ \\ \underline{F}_i \end{bmatrix} = n_i$ and $\underline{y}_i$ are r_i-vectors with rank $\underline{G}_i = r_i$. $\underline{A}$, $\underline{B}$, $\underline{C}$, $\underline{D}$, $\underline{E}_i$, $\underline{F}_i$, $\underline{G}_i$ are specified constant matrices of appropriate dimensions. The elements of the vectors $\underline{\omega}_i$ and $\underline{y}_i$, $i=1,2,\ldots,\Gamma$ are assumed to satisfy the differential equation:

$$\omega^{(p_i)} + q^i_{p_i}\,\omega^{(p_i-1)} + \ldots + q^i_2\,\dot{\omega} + q^i_1\,\omega = 0, \quad i=1,2,\ldots,\Gamma \tag{2}$$

---

* This work has been supported by the National Research Council of Canada under Grant No. A4396.

where $\mathrm{Re}(\tilde{\lambda}_j^i) \geq 0$, $j=1,2,\ldots,p_i$; $i=1,2,\ldots,\Gamma$ where $\tilde{\lambda}_j^i$ are the characteristic roots of (2). The initial conditions of (2) are assumed to be unknown for the elements of the vectors $\underline{\omega}_i$ and known for the elements of the vector $\underline{y}_i$. Without loss of generality, it is assumed that $\mathrm{rank}\underline{B}=m$, $\mathrm{rank}\underline{C}=r$ and that the following non-redunancy condition holds:

$$\mathrm{rank}\begin{bmatrix} \underline{E}_{r_k} & \underline{E}_{s_k} & \underline{E}_{t_k} & \cdots \\ \underline{F}_{r_k} & \underline{F}_{s_k} & \underline{F}_{t_k} & \cdots \end{bmatrix} = n_{r_k} + n_{s_k} + n_{t_k} + \ldots \; ; \; k=1,2,\ldots,k^* \qquad (3)$$

where $\bar{\lambda}_k$, $k=1,2,\ldots,k^*$ are the common roots for disturbances $\underline{\omega}_i$, $i=r_k, s_k, t_k, \ldots$ i.e. $\bar{\lambda}_k \varepsilon \{\tilde{\lambda}_j^i, j=1,2,\ldots,p_i\}$, $i=r_k, s_k, t_k, \ldots$, $k=1,2,\ldots,k^*$. It is desired to find a linear time-invariant controller for (1) assuming that (a) $\underline{x}$ and $\underline{y}$ are available for measurement or (b) $\underline{y}_m \triangleq \underline{C}_m\underline{x} + \underline{D}_m\underline{u} + \sum_{i=1}^{\Gamma} \underline{F}_i^m\underline{\omega}_i$ is only available for measurement, so that $\underline{e}(t)\to\underline{0}$ as $t\to\infty$, and such that the controlled system is controllable (or stabilizable) for the case when $\underline{\omega}_i$ is available for measurement and for the case when $\underline{\omega}_i$ is not available for measurement. In addition, it is desired to find a "robust controller" for the system, so that asymptotic tracking will take place independent of perturbations in the parameters of the system matrices $\underline{A}$, $\underline{B}$, $\underline{C}$, $\underline{D}$.

Davison has found a solution to the above problem, when $\Gamma=1$, rank $\underline{E}_1=n$, $\underline{G}_1=\underline{I}$, $\underline{D}=\underline{F}=\underline{0}$ in [1] for the case of a feedback controller and, when $\Gamma=1$, $\underline{G}_1=\underline{I}$, in [2] for the case of a feedforward controller. Wonham and Pearson [3] have considered a different formulation of basically the same problem with $\underline{D}=\underline{D}_m=\underline{0}$ and assuming that an observer is used by using a geometric approach. It is the purpose of this paper and its companion [4] to extend the results of [1] and [2] to include the more general formulation given in (1); in so doing, a complete characterization of controller structure is obtained which was not considered in [3]. Some other papers dealing with special cases of the above problem are given in [5] - [10].

II. MATHEMATICAL PRELIMINARIES

The following definitions are required in the development to follow:

Let λ_j^i, $j=1,2,\ldots\bar{p}^i$; $i=1,2,\ldots,\Gamma$ denote the characteristic roots of (2) defined such that λ_j^i is repeated q_{ij} times , $j=1,2,\ldots,\bar{p}^i$; $i=1,2,\ldots,\Gamma$ where

$$\sum_{j=1}^{\bar{p}^i} q_{ij} = p_i \,, \quad i=1,2,\ldots,\Gamma \tag{4}$$

Let $\underline{H}^i$ be a $mp_i \mathrm{x} mp_i$ matrix given by:

$$\underline{H}^i = \begin{bmatrix} \underline{N}_1^i \\ \underline{N}_2^i \\ \vdots \\ \underline{N}_{\bar{p}^i}^i \end{bmatrix} , \quad i=1,2,\ldots,\Gamma$$

where $\underline{N}_j^i$, $j=1,2,\ldots,\bar{p}^i$ is a $mq_{ij} \mathrm{x} mp_i$ matrix given by:

$$\underline{N}_j^i = \begin{bmatrix} \underline{I} & \lambda_j^i \, \underline{I} & \lambda_j^{i^2} \underline{I} & \lambda_j^{i^3} \underline{I} & \ldots & \lambda_j^{i^{p_i-1}} \underline{I} \\ \underline{0} & \underline{I} & \binom{2}{1}\lambda_j^i \, \underline{I} & \binom{3}{1}\lambda_j^{i^2} \underline{I} & \ldots & \binom{p_i-1}{1}\lambda_j^{i^{p_i-2}} \underline{I} \\ \underline{0} & \underline{0} & \underline{I} & \binom{3}{2}\lambda_j^i \, \underline{I} & \ldots & \binom{p_i-1}{2}\lambda_j^{i^{p_i-3}} \underline{I} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \underline{0} & \underline{0} & \underline{0} & \underline{0} \ldots \underline{I} & \ldots & \binom{p_i-1}{q_{ij}-1}\lambda_j^{i^{p_i-q_{ij}}} \underline{I} \end{bmatrix} , \quad j=1,2,\ldots,\bar{p}^i \tag{5}$$

Let $\underline{Q}_t^i$ be a $p_i t \mathrm{x} p_i t$ matrix given by:

$$\underline{Q}_t^i = \begin{bmatrix} \underline{0} & \underline{I} & \underline{0} & \ldots & \underline{0} \\ \underline{0} & \underline{0} & \underline{I} & \ldots & \underline{0} \\ \vdots & \vdots & \vdots & & \vdots \\ -q_1^i \, \underline{I} & -q_2^i \, \underline{I} & -q_3^i \, \underline{I} & \ldots & -q_{p_i}^i \, \underline{I} \end{bmatrix} , \quad i=1,2,\ldots,\Gamma \tag{6}$$

Let $\underline{C}_t^i$ be a $t \mathrm{x} p_i t$ matrix given by:

$$\underline{C}_t^i = (\underline{I}, \underline{0}, \underline{0}, \dots, \underline{0}) \quad , \quad i=1,2,\dots,\Gamma \tag{7}$$

Definitions

Assume that $\underline{y}_m \overset{\Delta}{=} \underline{C}_m \underline{x} + \underline{D}_m \underline{u} + \sum_{i=1}^{\Gamma} \underline{F}_i^m \underline{\omega}_i$ is only available for measurement, then: *a linear, time-invariant, feedforward control system* is one in which control applied to the plant is of the form:

$$\underline{u} = \underline{K}\hat{\underline{x}} + \sum_{i=1}^{\Gamma} (\underline{G}_i^* \underline{\xi}_i^* + \underline{G}_i \underline{\xi}_i)$$

where $\hat{\underline{x}}$, $\underline{\xi}_i^*$, $\underline{\xi}_i$ is the output of a linear time invariant system with inputs $\underline{y}_m$, $\underline{y}_i$, $\underline{\omega}_i$ (assumed measurable). $\underline{K}$, $\underline{G}_i^*$, $\underline{G}_i$ are constant gain matrices and $\underline{K}$ is found to stabilize $\underline{A}+\underline{B}\underline{K}$ so that it has an appropriate set of eigenvalues. *A linear, time-invariant, feedback control system* is one in which control applied to the plant is of the form: $\underline{u} = \underline{G}\,\underline{\xi}$ where $\underline{\xi}$ is the output of a linear time invariant system with inputs $\underline{y}_m$ and $\underline{y}_i$ $i=1,2,\dots,\Gamma$ and $\underline{G}$ is a constant gain matrix. Given the system (1), assume a feedback controller $\dot{\underline{\eta}} = \underline{\Lambda}_2\underline{\eta} + \underline{\Lambda}_1\underline{x}$, $\underline{u} = \underline{K}_1\underline{x} + \underline{K}_2\underline{\eta}$ has been applied; then the *controlled system is said to be controllable* if the pair

$$\left\{ \begin{bmatrix} \underline{A} & \underline{0} \\ \underline{\Lambda}_1 & \underline{\Lambda}_2 \end{bmatrix}, \begin{bmatrix} \underline{B} \\ \underline{0} \end{bmatrix} \right\}$$ is controllable.

III. FEEDFORWARD AND FEEDBACK CONTROL RESULTS*

It will initially be assumed that $\underline{x}$ and $\underline{y}$ are available for measurement in all control configurations. The case when $\underline{y}_m$ is only available for measurement will be dealt with in part II of the paper [4].

The following theorem is the main result of this paper:

Theorem 1

A necessary and sufficient condition that there exists a linear, time-invariant controller for (1) (either feedback or feedforward) such that $\underline{e}(t) \to \underline{0}$ as $t \to \infty$, for all disturbances $\underline{\omega}_i$, $i=1,2,\dots,\Gamma$ and all reference outputs $\underline{y}_i$, $i=1,2,\dots,\Gamma$ described by (2), and such that the controlled system is controllable (or stabilizable) is that the following two conditions both hold:

(a) $(\underline{A},\underline{B})$ be controllable (or stabilizable)

---

* All proofs are omitted in the paper; they may be found in Davison E.J. "The Feedforward and Feedback Control of a General Servo Problem - Part I", Department of Electrical Engineering, University of Toronto, Control System Report No. 7305 April 1973.

(b)
$$\text{rank}\begin{bmatrix} A-\lambda_j^i I & B \\ C & D \end{bmatrix} = \text{rank}\begin{bmatrix} A-\lambda_j^i I & B & E_i & 0 \\ C & D & F_i & G_i \end{bmatrix}$$

$$\text{rank}\begin{bmatrix} (A-\lambda_j^i I)^2 & (A-\lambda_j^i I)B & B \\ C & D & 0 \end{bmatrix} = \text{rank}\begin{bmatrix} (A-\lambda_j^i I)^2 & (A-\lambda_j^i I)B & B & E_i \\ C & D & 0 & 0 \end{bmatrix}$$

$$\text{rank}\begin{bmatrix} (A-\lambda_j^i I)^3 & (A-\lambda_j^i I)^2 B & (A-\lambda_j^i I)B & B \\ C & D & 0 & 0 \end{bmatrix}$$

$$= \text{rank}\begin{bmatrix} (A-\lambda_j^i I)^3 & (A-\lambda_j^i I)^2 B & (A-\lambda_j^i I)B & B & E_i \\ C & D & 0 & 0 & 0 \end{bmatrix}$$

$$\vdots$$

$$\text{rank}\begin{bmatrix} (A-\lambda_j^i I)^{q_{ij}} & (A-\lambda_j^i I)^{q_{ij}-1}B & (A-\lambda_j^i I)^{q_{ij}-2}B & \cdots & (A-\lambda_j^i I)B & B \\ C & D & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$= \text{rank}\begin{bmatrix} (A-\lambda_j^i I)^{q_{ij}} & (A-\lambda_j^i I)^{q_{ij}-1}B & (A-\lambda_j^i I)^{q_{ij}-2}B & \cdots & (A-\lambda_j^i I)B & B & E_i \\ C & D & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}$$

$$j=1,2,\ldots,\bar{p}^i \; ; \; i=1,2,\ldots,\Gamma \qquad (8)$$

Discussion of Theorem 1

1. For *tracking problems* with or without disturbances present, e.g. when rank $G_i = r$, $i=1,2,\ldots,\Gamma$, condition (b) of Theorem 1 becomes simply that:

$$\text{rank}\begin{bmatrix} A-\lambda_j^i I & B \\ C & D \end{bmatrix} = n+r \quad , \quad j=1,2,\ldots,\bar{p}^i \; ; \; i=1,2,\ldots,\Gamma \qquad (9)$$

This agrees with the results of [1], [2] for the case that $\Gamma=1$ (after some simple manipulations). Note that there is no solution to the problem if $r > m$ in this case.

The following feedforward controller will accomplish the objectives of Theorem 1.

Feedforward Controller

$$\underline{u} = \underline{K}\underline{x} + \sum_{i=1}^{\Gamma} \left\{ (\underline{G}_0^i, \underline{G}_1^i, \ldots, \underline{G}_{p_i-1}^i)\hat{\underline{\eta}}_i + (\underline{G}_0^{i*}, \underline{G}_1^{i*}, \ldots, \underline{G}_{p_i-1}^i)\hat{\underline{\eta}}_i^{*} \right\} \tag{10}$$

where $\hat{\underline{\eta}}_i$ is the output of an observer with input $\underline{\omega}_i$ which estimates the state $\underline{\eta}_i$ of the observable system:

$$\dot{\underline{\eta}}_i = \underline{Q}^i_{n_i}\underline{\eta}_i \quad , \quad \underline{\omega}_i = \underline{C}^i_{n_i}\underline{\eta}_i \quad , \quad i=1,2,\ldots,\Gamma \tag{11}$$

and $\hat{\underline{\eta}}_i^{*}$ is the output of an observer with input $\underline{y}_i$ which estimates the state $\underline{\eta}_i^{*}$ of the observable system:

$$\dot{\underline{\eta}}_i^{*} = \underline{Q}^i_{r_i}\underline{\eta}_i^{*} \quad , \quad \underline{y}_i = \underline{C}^i_{r_i}\underline{\eta}_i^{*} \quad , \quad i=1,2,\ldots,\Gamma \tag{12}$$

where $\underline{K}$ is a feedback gain matrix found so that $(\underline{A} + \underline{B}\underline{K})$ is stable with suitable fast dynamics using e.g. pole assignment or linear optimal control theory, and the $m \times n_i$ *disturbance gain matrices* $\underline{G}_j^i$, $j=0,1,\ldots,p_i-1$, $i=1,2,\ldots,\Gamma$ are given as follows:

$$\begin{bmatrix} \underline{G}_0^i \\ \underline{G}_1^i \\ \cdot \\ \cdot \\ \cdot \\ \underline{G}_{p_i-1}^i \end{bmatrix} = \{H^i\}^{-1} \begin{bmatrix} \underline{\Lambda}_1^i \\ \underline{\Lambda}_2^i \\ \cdot \\ \cdot \\ \cdot \\ \underline{\Lambda}_{\bar{p}^i}^i \end{bmatrix} \quad , \ i=1,2,\ldots,\Gamma \tag{13}$$

where $\underline{\Lambda}_k^i$, $k=1,2,\ldots,\bar{p}^i$ is a $mq_{ik} \times n_i$ matrix given as follows:

$$\underline{\Lambda}_k^i = \begin{bmatrix} {}^i\underline{\delta}_k^1 \\ {}^i\underline{\delta}_k^2 \\ \cdot \\ \cdot \\ \cdot \\ {}^i\underline{\delta}_k^{q_{ik}} \end{bmatrix} \quad , \ k=1,2,\ldots,\bar{p}^i \ ; \ i=1,2,\ldots,\Gamma \tag{14}$$

where

$$
\begin{aligned}
i\underline{\delta}_k^1 &= -(i\underline{\mathcal{E}}_k^1 + \underline{D})^{\dagger}(i\underline{\mathcal{D}}_k^1 + \underline{F}_i) \\
i\underline{\delta}_k^2 &= (i\underline{\mathcal{E}}_k^1 + \underline{D})^{\dagger}(i\underline{\mathcal{D}}_k^2 + i\underline{\mathcal{E}}_k^2 i\underline{\delta}_k^1) \\
i\underline{\delta}_k^3 &= (i\underline{\mathcal{E}}_k^1 + \underline{D})^{\dagger}(-i\underline{\mathcal{D}}_k^3 - i\underline{\mathcal{E}}_k^3 i\underline{\delta}_k^1 + i\underline{\mathcal{E}}_k^2 i\underline{\delta}_k^2) \\
i\underline{\delta}_k^4 &= (i\underline{\mathcal{E}}_k^1 + \underline{D})^{\dagger}(i\underline{\mathcal{D}}_k^4 + i\underline{\mathcal{E}}_k^4 i\underline{\delta}_k^1 - i\underline{\mathcal{E}}_k^3 i\underline{\delta}_k^2 + i\underline{\mathcal{E}}_k^2 i\underline{\delta}_k^3) \\
&\;\;\vdots \\
i\underline{\delta}_k^{q_{ik}} &= (i\underline{\mathcal{E}}_k^1 + \underline{D})^{\dagger}\Big[(-)^{q_{ik}} i\underline{\mathcal{D}}_k^{q_{ik}} + (-)^{q_{ik}} i\underline{\mathcal{E}}_k^{q_{ik}} i\underline{\delta}_k^1 + (-)^{q_{ik}-1} i\underline{\mathcal{E}}_k^{q_{ik}-1} i\underline{\delta}_k^2 + \ldots \\
&\qquad \ldots + i\underline{\mathcal{E}}_k^2 i\underline{\delta}_k^{q_{ik}-1}\Big] \\
&\qquad k=1,2,\ldots,\bar{p}^i \;;\quad i=1,2,\ldots,\Gamma
\end{aligned}
\tag{15}
$$

where

$$
\left.
\begin{aligned}
i\underline{\mathcal{E}}_k^{\theta} &= (\underline{C}+\underline{D}\underline{K})(\lambda_k^i\underline{I}-\underline{A}-\underline{B}\underline{K})^{-\theta}\underline{B} \\
i\underline{\mathcal{D}}_k^{\theta} &= (\underline{C}+\underline{D}\underline{K})(\lambda_k^i\underline{I}-\underline{A}-\underline{B}\underline{K})^{-\theta}\underline{E}_i
\end{aligned}
\right\} \;;\; \theta=1,2,\ldots,q_{ik} \;;\; k=1,2,\ldots,\bar{p}^i
\tag{16}
$$

where $(i\underline{\mathcal{E}}_k^1 + \underline{D})^{\dagger}$ denotes the pseudo-inverse of $(i\underline{\mathcal{E}}_k^1 + \underline{D})$.

The $m \times r_i$ *tracking gain matrices* $\underline{G}_j^{*i}$, $j=0,1,\ldots,p_i-1$; $i=1,2,\ldots,\Gamma$ are equal to $\underline{G}_j^i$, $j=0,1,\ldots,p_i-1$; $i=1,2,\ldots,\Gamma$ respectively with $\underline{E}_i$ being replaced by $\underline{0}$ and $\underline{F}_i$ being replaced by $-\underline{G}_i$.

Remarks on Feedforward Controller

2. The tracking and disturbance gain matrices given by (13) are unique if and only if rank $\begin{bmatrix} \underline{A}-\lambda_j^i\underline{I} & \underline{B} \\ \underline{C} & \underline{D} \end{bmatrix} = n+m$, $j=1,2,\ldots,\bar{p}^i$; $i=1,2,\ldots,\Gamma$ and in this case

$$
(i\underline{\mathcal{E}}_k^1 + \underline{D})^{\dagger} = \left\{(i\underline{\mathcal{E}}_k^1 + \underline{D})'(i\underline{\mathcal{E}}_k^1 + \underline{D})\right\}^{-1}(i\underline{\mathcal{E}}_k^1 + \underline{D})'
\tag{17}
$$

In case this condition is not satisfied, the gain matrices of the feedforward controller will have minimum norm.

3. Standard observer construction [11] is used to implement the tracking and/or disturbance observers. It should be noted that a particular

simple observer implementation is the following:

$$\underline{u} = \underline{K}\underline{x} + \sum_{i=1}^{\Gamma} \left\{ (\underline{G}_0^i + \underline{G}_1^i s + \ldots + \underline{G}_{p_i-1}^i s^{p_i-1})\underline{\omega}_i + (\underline{G}_0^{i*} + \underline{G}_0^{i*} s + \ldots + \underline{G}_{p_i-1}^{i*} s^{p_i-1})\underline{y}_i \right\} \tag{18}$$

where s denotes the operator $\frac{d}{dt}(\cdot)$, which however because of noise considerations, would generally be only practical to apply with $p_i \leq 2$.

4. It is not essential to have state feedback in this controller (or any feedback at all if $\underline{A}$ is stable), i.e. $\underline{K}$ is found solely to stabilize and "speed up the dynamics" of $\underline{A} + \underline{B}\underline{K}$; this has extremely important engineering applications.

In case $\underline{\omega}_i$ is not available for measurement, the following feedback control system must be used:

Feedback Controller

$$\underline{u} = \sum_{i=1}^{\Gamma} (\underline{G}_0^{*i}, \underline{G}_1^{*i}, \ldots, \underline{G}_{p_i-1}^{*i})\hat{\underline{\eta}}_i^* + \hat{\underline{u}} \tag{19}$$

where $\hat{\underline{u}} = \left\{ \underline{K} , (\underline{G}_0^1, \ldots, \underline{G}_{p_1-1}^1) , \ldots, (\underline{G}_0^{\Gamma}, \ldots, \underline{G}_{p_\Gamma-1}^{\Gamma}) \right\} \hat{\underline{\xi}}$, where $\hat{\underline{\xi}}$ is the output of an observer with input $\begin{bmatrix} \underline{x} - \tilde{\underline{x}} \\ \underline{e} \end{bmatrix}$ which estimates the state $\underline{\xi} \triangleq (\underline{x}', \underline{\eta}_1', \underline{\eta}_2', \ldots, \underline{\eta}_\Gamma')'$ of the observable system:

$$\dot{\underline{\xi}} = \underline{A}_{aug}\underline{\xi} + \underline{B}_{aug}\hat{\underline{u}} \quad , \quad \begin{bmatrix} \underline{x} - \tilde{\underline{x}} \\ \underline{e} \end{bmatrix} = \underline{C}_{aug}\underline{\xi} + \underline{D}_{aug}\hat{\underline{u}} \tag{20}$$

where $\tilde{\underline{x}}$ is the output of the compensator:

$$\dot{\underline{\sigma}} = (\underline{A}+\underline{B}\underline{K})\underline{\sigma} + \underline{B} \sum_{i=1}^{\Gamma} (\underline{G}_0^{*i}, \ldots, \underline{G}_{p_i-1}^{*i})\hat{\underline{\eta}}_i^* \tag{21}$$

$$\tilde{\underline{x}} = \underline{\sigma}$$

where $\underline{A}_{aug}$, $\underline{B}_{aug}$, $\underline{C}_{aug}$, $\underline{D}_{aug}$ are given as follows:

$$\underline{A}_{aug} = \begin{bmatrix} \underline{A} & \underline{E}_1(\underline{I},\underline{0}) & \underline{E}_2(\underline{I},\underline{0}) & \ldots & \underline{E}_\Gamma(\underline{I},\underline{0}) \\ \underline{0} & \underline{Q}_{n_1}^1 & \underline{0} & \ldots & \underline{0} \\ \underline{0} & \underline{0} & \underline{Q}_{n_2}^2 & \ldots & \underline{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \underline{0} & \underline{0} & \underline{0} & \ldots & \underline{Q}_{n_\Gamma}^\Gamma \end{bmatrix} \tag{22}$$

$$\underline{B}_{aug}' = (\underline{B}' , \underline{0} , \underline{0} , \ldots , \underline{0}) \tag{23}$$

$$C_{aug} = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ C & F_1(I,0) & F_2(I,0) & \cdots & F_\Gamma(I,0) \end{bmatrix} \tag{24}$$

$$D_{aug} = \begin{bmatrix} 0 \\ D \end{bmatrix} \tag{25}$$

and all other quantities are defined as before in the feedforward controller.

Discussion of Theorem 1

It may immediately be observed from Theorem 1 that there is almost always a solution to the servo problem for any (A, B, C, D) provided rank $B \geq$ rank C, i.e. the class of (A, B, C, D) which does not satisfy the conditions of Theorem 1 in this case lies on a hypersurface in the parameter space of (A, B, C, D). (See [12] for a definition and properties of a hypersurface). If rank $B <$ rank C, there is almost never a solution to the problem i.e. (A, B, C, D) must lie on a hypersurface in the parameter space of (A, B, C, D) in order for there to be a solution; in particular, if rank $B <$ rank C there is never a solution for the tracking problem and for regulation problems in which rank $E_i = n$, $F_i = 0$ or rank $F_i = r$, $E_i = 0$.

Thus there is almost never a solution to the servo problem if there are more outputs to be controlled than inputs.

The frequency domain interpretation of condition (b) of Theorem 1 is interesting; it means that either no *transmission zeros* of the system $\dot{x} = Ax + Bu$, $y = Cx + Du$ coincide with the characteristic roots λ_j^i of the differential equation describing the class of reference outputs or/and disturbance inputs (see [13] for a definition of transmission zeros) i.e.

$$\text{rank} \begin{bmatrix} A-\lambda_j^i I & B \\ C & D \end{bmatrix} = n+r, \quad j=1,2,\ldots,p_i; \quad i=1,2,\ldots,\Gamma, \tag{26}$$

or if a transmission zero of the system does coincide with a root λ_j^i, then the disturbance-tracking vectors (E_i, F_i, G_i) must be a specific linear combination of the system's parameters (A, B, C, D) in the way as given in Theorem 1. If the transmission zeros of (A, B, C, D) coincide with the roots λ_j^i, then tracking cannot take place.

IV. NUMERICAL EXAMPLE

Consider the controllable system:

$$\dot{x} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} u + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \omega \tag{27}$$

$$y = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} x + u$$

where ω is a sinusoidal disturbance with frequency Ω . It is desired to find a controller so that $y \to 0,\ t \to \infty$. In this case from Theorem 1, there is a solution if and only if:

$$\text{rank} \begin{bmatrix} -j\Omega & 0 & -1 & 1 \\ 0 & -j\Omega & 0 & 1 \\ 1 & 0 & -j\Omega & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} = \text{rank} \begin{bmatrix} -j\Omega & 0 & -1 & 1 & 1 \\ 0 & -j\Omega & 0 & 1 & 0 \\ 1 & 0 & -j\Omega & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{28}$$

i.e. provided $\Omega \neq \pm 1$. Hence there is a solution for all Ω except $\Omega = \pm 1$.

REFERENCES

[1] Davison E.J., "The Output Control of Linear Time-Invariant Multivariable Systems with Unmeasurable Arbitrary Disturbances", IEEE Trans. on Automatic Control, Vol. AC-17, No.5, Oct. 1972, pp. 621-630.

[2] Davison, E.J., "The Feedforward Control of Linear Multivariable Time-Invariant Systems", to appear in Automatica.

[3] Wonham W.M., Pearson J.B., "Regulation and INternal Stabilization in Linear Multivariable Systems", Dept. of Electrical Engineering, University of Toronto, Control System Report No. 7212, Aug. 1972.

[4] Davison E.J., "The Feedforward and Feedback Control of a General Servomechanism Problem - Part II", Dept. of Electrical Engineering, Control System Report 7306, University of Toronto, April 1973.

[5] Davison E.J., Smith H.W., "Pole Assignment in Linear Time-Invariant Multivariable Systems with Constant Disturbances", Automatica, Vol. 7, 1971, pp. 489-498.

[6] Smith, H.W., Davison E.J., "The Design of Industrial Regulators: Integral Feedback and Feedforward Control", Proc IEE, Vol. 119, No. 8, Aug. 1972, pp. 1210-1216.

[7] Johnson C.D., "Accommodation of External Disturbances in Linear Regulator and Servomechanism Problems", IEEE Trans. on Automatic Control, Vol. AC-16, 1971, pp. 635-644.

[8] Bhattachary S.P., Pearson J.B., "Error Systems and the Servomechanism Problem", International J. of Control Vol.15,No.5,1972,pp.1041-1062.

[9] Young P.C., Willems J.C., "An Approach to the Multivariable Servomechanism Problem", International J. of Control, Vol. 15, No.5, 1972, pp. 961-979.

[10] Sandell N., Athans M., "On Type L' Multivariable Linear Systems", Automatica, Vol. 9, No. 1, 1973, pp. 131-136.

[11] Luenberger D.G., "Observers for Multivariable Systems", IEEE Trans. on Automatic Control, Vol. AC-11, No. 2, April 1966, pp. 190-197.

[12] Davison E.J., Wang S.H., "Properties of Linear Multivariable Systems Subject to Arbitrary Output and State Feedback", IEEE Trans. on Automatic Control, Vol. AC-18, No. 1, 1973, pp. 24-32.

[13] Wang, S.H., Davison E.J., "Canonical Forms of Linear Multivariable Systems", Dept. of Electrical Engineering, University of Toronto, Control System Report No. 7203, April 1972, presented at the 6th Asilomar Conference on Circuits and Systems 1972.

THE FEEDFORWARD AND FEEDBACK CONTROL OF A GENERAL SERVOMECHANISM PROBLEM

Part II

EDWARD J. DAVISON*
Department of Electrical Engineering
University of Toronto, Canada

ABSTRACT

In Part I [1] of the paper necessary and sufficient conditions were found for a linear time-invariant feedback or feedforward controller to exist for a general linear multivariable system subject to a general class of unmeasurable or/and measurable arbitrary disturbances, so that asymptotic regulation takes place and so that the controlled system is controllable (or stabilizable). This part of the paper gives a sensitivity analysis which characterizes the class of "robust" controllers which allows asymptotic tracking to take place independent of perturbations in the parameters of the system and obtains a minimal order "robust" controller. In addition an extension of the previous results is made to include the case when state or output feedback is not allowed.

I. INTRODUCTION

Consider the following linear time-invariant system:

$$\dot{\underline{x}} = \underline{A}\underline{x} + \underline{B}\underline{u} + \sum_{i=1}^{\Gamma} (\underline{E}_i, \underline{0}) \begin{bmatrix} \underline{\omega}_i \\ \underline{y}_i \end{bmatrix} \tag{1}$$

$$\underline{e} = \underline{C}\underline{x} + \underline{D}\underline{u} + \sum_{i=1}^{\Gamma} (\underline{F}_i, -\underline{G}_i) \begin{bmatrix} \underline{\omega}_i \\ \underline{y}_i \end{bmatrix}$$

where $\underline{x}$ is the n-state vector, $\underline{u}$ is the m-control vector, $\underline{e}$ the r-error vector, $\underline{y} \triangleq \underline{C}\underline{x} + \underline{D}\underline{u} + \sum_{i=1}^{\Gamma} \underline{F}_i\underline{\omega}_i$ the r-output vector and $\underline{y}_{ref} \triangleq \sum_{i=1}^{\Gamma} \underline{G}_i\underline{y}_i$ is the desired specified asymptotic (as $t\to\infty$) r-output vector. $\underline{\omega}_i$ are n_i disturbance vectors which may or may not be measurable. The elements of the vectors $\underline{\omega}_i$ and $\underline{y}_i$, $i=1,2,\dots,\Gamma$ are assumed to satisfy the differential equation:

$$\omega^{(p_i)} + q^i_{p_i}\, \omega^{(p_i-1)} + \dots + q^i_2\dot{\omega} + q^i_1\omega = 0 \quad , \quad i=1,2,\dots,\Gamma \tag{2}$$

where $\mathrm{Re}(\tilde{\lambda}^i_j) \geq 0$, $j=1,2,\dots,p_i$; $i=1,2,\dots,\Gamma$ where $\tilde{\lambda}^i_j$ are the characteristic roots of (2). The initial conditions of (2) are assumed to be unknown for the elements of the vectors $\underline{\omega}_i$ and known for the elements

---

* This work has been supported by the National Research Council of Canada under Grant No. A4396.

of the vector $\underline{y}_i$. It is assumed that (1) satisfies certain non-redunancy conditions as given in Part I [1] of the paper. For the sake of brevity these conditions will not be repeated here. Also the same notation and definitions as used in Part I [1] of the paper will be assumed and for the sake of brevity will not be repeated here.

It is desired to find a linear time-invariant controller for (1) assuming that (a) $\underline{x}$ and $\underline{y}$ are available for measurement or (b) $\underline{y}_m = \underline{C}_m \underline{x} + \underline{D}_m \underline{u} + \sum_{i=1}^{\Gamma} \underline{F}_i^m \underline{\omega}_i$ is only available for measurement, so that $\underline{e}(t) \rightarrow \underline{0}$ as $t \rightarrow \infty$, and such that the controlled system is controllable (or stabilizable) for the case when $\underline{\omega}_i$ is available and for the case when $\underline{\omega}_i$ is not available for measurement. Part I [1] of the paper obtained necessary and sufficient conditions for this to be done for the case of (a) and obtained explicit feedforward and feedback controllers. These controllers however suffer from the disadvantage that regulation is in general not achieved if any of the system's parameters $(\underline{A}, \underline{B}, \underline{C}, \underline{D})$ or feedback gains are perturbed. This part of the paper deals with the case of robust control systems in which asymptotic regulation is assured even if the system's parameters are perturbed, and with the general case when $\underline{y}_m$ is only available for measurement.

II. ROBUST CONTROL RESULTS*

The following definitions will be made:

Definition

Given a real nxm matrix $\underline{M}$, let $\underline{M} \varepsilon \Omega_\varepsilon$,where $\underline{M} = \{m_{ij}\}$, $i=1,2,\ldots,n$; $j=1,2,\ldots,m$ denote $m_{ij} \varepsilon \Omega_\varepsilon$, $i=1,2,\ldots,n$; $j=1,2,\ldots,m$ where

$$\Omega_\varepsilon = \left\{ m_{ij} \,\middle|\, |m_{ij}| < \varepsilon \right\} \tag{3}$$

Definition

Given the system (1), suppose that there exists a controller so that asymptotic regulation takes place and so that the controlled system is stable. Let the plant parameters $\underline{A}$ and $\underline{B}$ now be perturbed i.e. $\underline{A} \rightarrow \underline{A} + \delta\underline{A}$, $\underline{B} \rightarrow \underline{B} + \delta\underline{B}$ where $\delta\underline{A}\varepsilon\Omega_\varepsilon$, $\delta\underline{B}\varepsilon\Omega_\varepsilon$. Then if there exists an $\varepsilon>0$ so that the controlled system is still stable and such that asymptotic regulation still takes place for $\forall\delta\underline{A}\varepsilon\Omega_\varepsilon$, $\forall\delta\underline{B}\varepsilon\Omega_\varepsilon$, the controller is said to be a *robust controller*.

Definition

Let λ_i^*, δ_i; $i=1,2,\ldots,\phi$ be defined as follows. Consider the Γ

---

* All proofs are omitted; they may be found in Davison E.J., "The Feedforward and Feedback Control of a General Servomechanism Problem - Part II", Department of Electrical Engineering, University of Toronto, Control System Report 7306 April 1973.

sets of distinct roots $\{\lambda_j^i, j=1,2,\ldots,\bar{p}^i\}$, $i=1,2,\ldots,\Gamma$. Suppose there exists a root $\lambda_k^r, \lambda_\ell^s, \lambda_m^t,\ldots$ which is common to sets $r,s,t,\ldots$ and suppose that $q_{rk} \geq \max\{q_{s\ell}, q_{tm},\ldots\}$; then retain the root λ_k^r in the r^{th} set and eliminate the roots $\lambda_\ell^s, \lambda_m^t,\ldots$ occurring in the s^{th}, t^{th},... sets respectively. Continue this process until there are no common roots occurring in the non-empty sets which remain. Let the collection of all remaining roots in all the non-empty sets be given by $\left\{\lambda_{b_1}^{a_1}, \lambda_{b_2}^{a_2},\ldots,\lambda_{b_z}^{a_z}\right\}$ where the integers $a_1, a_2,\ldots,a_z$; $b_1, b_2,\ldots,b_z$ are not necessarily distinct. Then λ_j^*, $j=1,2,\ldots,\phi$ is defined to be as follows:

$$\lambda_j^*,\ j=1,2,\ldots,\phi = \left\{\underbrace{\lambda_{b_1}^{a_1}, \lambda_{b_1}^{a_1},\ldots,\lambda_{b_1}^{a_1}}_{\text{repeated } q_{a_1,b_1} \text{ times}};\ \underbrace{\lambda_{b_2}^{a_2}, \lambda_{b_2}^{a_2},\ldots,\lambda_{b_2}^{a_2}}_{\text{repeated } q_{a_2,b_2} \text{ times}};\ \ldots;\ \underbrace{\lambda_{b_z}^{a_z}, \lambda_{b_z}^{a_z},\ldots,\lambda_{b_z}^{a_z}}_{\text{repeated } q_{a_z,b_z} \text{ times}}\right\} \tag{4}$$

and the expansion of $\prod_{j=1}^{\phi}(\lambda-\lambda_j^*)$ given by:

$$\prod_{j=1}^{\phi}(\lambda-\lambda_j^*) = \lambda^{\phi} + \delta_{\phi}\lambda^{\phi-1} + \delta_{\phi-1}\lambda^{\phi-2} + \ldots + \delta_2\lambda + \delta_1 \tag{5}$$

defines the coefficients δ_i, $i=1,2,\ldots,\phi$.

The following result will now be proved. It is assumed that the state $\underline{x}$ is available for measurement.

Theorem 1

A necessary and sufficient condition that there exists a robust linear, time-invariant controller for (1) such that $\underline{e}(t)\rightarrow\underline{0}$ as $t\rightarrow\infty$ for all measurable or unmeasurable disturbances $\underline{\omega}_i$ and all specified outputs $\underline{y}_i$ described by (2), and such that the controlled system is controllable (or stabilizable) is that the following two conditions both hold:

(a) $(\underline{A},\underline{B})$ be controllable (or stabilizable)

(b) $$\text{rank}\begin{bmatrix} \underline{A}-\lambda_j^i I & \underline{B} \\ \underline{C} & \underline{D} \end{bmatrix} = n+r\ ,\ j=1,2,\ldots,\bar{p}^i\ ;\ i=1,2,\ldots,\Gamma \tag{6}$$

In case these two conditions hold, the robust controller must be a feedback controller of order at least $r\phi$ with $\underline{e}(t)$ as an input.

A *robust controller* of minimum order $(r\phi)$ which satisfies the requirements of the above theorem is given as follows:

$$\underline{u} = \underline{K}_0\underline{x} + \sum_{i=1}^{r} \underline{K}_i\underline{\eta}_i \tag{7}$$

where

$$\dot{\underline{\eta}}_i = \underline{C}\,\underline{\eta}_i + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} e_i \quad , \quad i=1,2,\ldots,r \tag{8}$$

where

$$\underline{e} = (e_1, e_2, \ldots, e_r)' \quad \text{and}$$

$$\underline{C} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -\delta_1 & -\delta_2 & -\delta_3 & & -\delta_\phi \end{bmatrix} \tag{9}$$

The initial conditions $\underline{\eta}_i(0)$, $i=1,2,\ldots,r$ are unspecified and the gain matrices $\underline{K}_0$, $\underline{K}_i$, $i=1,2,\ldots,r$ are found so that $\{\underline{\mathcal{A}} + \underline{\mathcal{B}}(\underline{K}_0,\underline{K}_1,\ldots,\underline{K}_r)\}$ has stable specified dynamic behaviour, where the pair $(\underline{\mathcal{A}}, \underline{\mathcal{B}})$ is controllable (or stabilizable) and is given as follows:

$\underline{\mathcal{A}}$ is a $(n+r\phi)\mathrm{x}(n+r\phi)$ matrix and $\underline{\mathcal{B}}$ is a $(n+r\phi)\mathrm{x}m$ matrix:

$$\underline{\mathcal{A}} = \begin{bmatrix} \underline{A} & \underline{0} & \underline{0} & \ldots & \underline{0} \\ \begin{pmatrix} \underline{0} \\ \underline{c}_1' \end{pmatrix} & \underline{C} & \underline{0} & \ldots & \underline{0} \\ \begin{pmatrix} \underline{0} \\ \underline{c}_2' \end{pmatrix} & \underline{0} & \underline{C} & \ldots & \underline{0} \\ \vdots & \vdots & \vdots & & \\ \begin{pmatrix} \underline{0} \\ \underline{c}_r' \end{pmatrix} & \underline{0} & \underline{0} & \ldots & \underline{C} \end{bmatrix} , \quad \underline{\mathcal{B}} = \begin{bmatrix} \underline{B} \\ \begin{pmatrix} \underline{0} \\ \underline{d}_1' \end{pmatrix} \\ \begin{pmatrix} \underline{0} \\ \underline{d}_2' \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \underline{0} \\ \underline{d}_r' \end{pmatrix} \end{bmatrix} \tag{10}$$

where $\underline{C}' = (\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_r)$.

Remarks

1. It will be shown that any robust controller must be a feedback controller of order at least $r\phi$ with error as an input (with other inputs possibly). This is an extremely important observation; it justifies the

widespread use of error feedback in practical control configurations.

*2. The above controller (7), (8) has the property that it brings about asymptotic regulation for all disturbances, for *any finite changes* (and not just arbitrarily small perturbations) in the system's parameters $\underline{A}$, $\underline{B}$, $\underline{C}$, $\underline{D}$ or feedback gain matrices $\underline{K}_0$, $\underline{K}_i$, $i=1,2,\ldots,r$ and *for any changes in the order* of the assumed mathematical model describing the system provided only that the controlled system remains stable.

3. Note that Theorem 1 holds for almost all $(\underline{A}, \underline{B}, \underline{C}, \underline{D})$ in which rank $\underline{B} \geq$ rank $\underline{C}$, i.e. the class of $(\underline{A}, \underline{B}, \underline{C}, \underline{D})$ in which Theorem 1 does not hold, lies on a hypersurface in the parameter space of $(\underline{A}, \underline{B}, \underline{C}, \underline{D})$ [2]. If rank $\underline{B} <$ rank $\underline{C}$ there is no solution to the problem.

REMARK ON CHARACTERIZATION OF ALL MINIMAL ORDER ROBUST CONTROLLERS

The minimal order controller of Theorem 1 is not unique. The most general minimal order robust controller has the following structure:

$$\underline{u} = \underline{K}_0\underline{x} + \underline{K}\underline{\xi} \tag{11}$$

where

$$\dot{\underline{\xi}} = \underline{\tau}\begin{bmatrix} \underline{C} & \underline{0} & \cdots & \underline{0}\\ \underline{0} & \underline{C} & \cdots & \underline{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{0} & \underline{0} & & \underline{C}\end{bmatrix}\underline{\tau}^{-1}\underline{\xi} + \underline{\tau}\begin{bmatrix} \begin{bmatrix}\theta_1\\ \vdots \\ \theta_\phi\end{bmatrix} & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \begin{bmatrix}\theta_{\phi+1}\\ \vdots\\ \theta_{2\phi}\end{bmatrix} & \cdots & \underline{0}\\ \vdots & \vdots & & \\ \underline{0} & \underline{0} & \cdots & \begin{bmatrix}\theta_{r\phi-\phi}\\ \vdots \\ \theta_{r\phi}\end{bmatrix}\end{bmatrix}\begin{bmatrix}e_1\\ e_2\\ \vdots\\ e_r\end{bmatrix}. \tag{12}$$

where $\underline{\tau}$ is an arbitrary nonsingular real matrix, $\underline{C}$ is defined in (9) and θ_i, $i=1,2,\ldots,r\phi$ are arbitrary real scalars chosen so that:

---

* Note that it has been assumed that the error $\underline{e} \triangleq (\underline{y}-\underline{y}_{ref})$ can be measured and that the elements of the controller dynamics $\delta_1,\ldots,\delta_\phi$ can be made precise - this is usually the case in practise, especially when the disturbance - reference outputs are polynomial or sinusoidal functions of time.

$$\left\{\underline{C}\,,\begin{bmatrix}\theta_1\\ \vdots\\ \theta_{\phi}\end{bmatrix}\right\},\ \left\{\underline{C}\,,\begin{bmatrix}\theta_{\phi+1}\\ \vdots\\ \theta_{2\phi}\end{bmatrix}\right\},\ \ldots,\ \left\{\underline{C}\,,\begin{bmatrix}\theta_{r\phi-\phi}\\ \vdots\\ \theta_{r\phi}\end{bmatrix}\right\} \tag{13}$$

are all controllable. It may be noted that the controller (11), (12) consists of an identical number of unstable compensators (with dynamics corresponding to the class of disturbance/reference outputs) connected to the error outputs of the system.

III. RESULTS FOR THE CASE WHEN $\underline{y}_m$ IS ONLY AVAILABLE FOR MEASUREMENT

In all the previous development, it has been assumed that the state $\underline{x}$ and output $\underline{y}$ have been available for measurement. This assumption will now be relaxed. It will be assumed that $\underline{y}_m \triangleq \underline{C}_m\underline{x} + \underline{D}_m\underline{u} + \sum_{i=1}^{\Gamma} \underline{F}_i^m \underline{\omega}_i$ is only available for measurement. (It is of course assumed that $\underline{y}_i$ is still available for measurement.) The following results are now obtained:

Disturbance Observation

Occasionally, it may be desirable to be able to estimate the unmeasurable disturbances $\underline{\omega}_i$, $i=1,2,\ldots,\Gamma$ of the system (1) from a measurement of $\underline{y}_m$, e.g. in multi-stage industrial processes where disturbances observed in the first stage can be used to generate feedforward controls for later stages. The following theorem gives conditions under which this can be done:

Definition

Let $\bar{\lambda}_k$, $k=1,2,\ldots k^*$ be a common root for the disturbances $\underline{\omega}_i$, $i=r_k,s_k,t_k,\ldots$; i.e. $\bar{\lambda}_k \varepsilon \left\{\lambda_j^i,\ j=1,2,\ldots,\bar{p}^i\right\}$, $i=r_k,s_k,t_k,\ldots$ for $k=1,2,\ldots,k^*$.

Theorem 2

Given the system (1) with measurement vector $\underline{y}_m$, then necessary and sufficient conditions that the disturbances $\underline{\omega}_i$, $i=1,2,\ldots,\Gamma$ can be estimated from $\underline{y}_m$ are that the following conditions both hold:

$$\text{(a)}\quad \text{rank}\begin{bmatrix}\underline{A}-\lambda_j^i\underline{I} & \underline{E}_i\\ \underline{C}_m & \underline{F}_i^m\end{bmatrix} = n+n_i\ ,\ j=1,2,\ldots,\bar{p}^i\ ;\ i=1,2,\ldots,\Gamma \tag{14}$$

$$\text{(b)}\quad \text{rank}\begin{bmatrix}\underline{A}-\bar{\lambda}_k\underline{I} & \underline{E}_{r_k} & \underline{E}_{s_k} & \underline{E}_{t_k} & \ldots\\ \underline{C}_m & \underline{F}_{r_k}^m & \underline{F}_{s_k}^m & \underline{F}_{t_k}^m & \ldots\end{bmatrix} = n+n_{r_k}+n_{s_k}+n_{t_k}+\ldots,\ k=1,2,\ldots k^* \tag{15}$$

Remarks

4. If it is desired to also estimate the state of the system, then the extra requirement of $(\underline{A},\underline{C}_m)$ being observable is also required.

5. It may be noted that the minimum number of independent measuring outputs $\dim(\underline{y}_m) \triangleq \operatorname{rank} \underline{C}_m$ required to estimate $\underline{\omega}_i$ is given as follows:

$$\dim(\underline{y}_m) = \max\left\{ n_i,\ i=1,2,\ldots,\Gamma\ ;\ n_{r_k} + n_{s_k} + n_{t_k} + \ldots,\ k=1,2,\ldots,k^* \right\}$$

Feedforward Results

Due to space limitations results on Feedforward Control are omitted.

Robust Control Results

In this case from Theorem 1, the robust controller must consist of error feedback i.e. there is no solution to the problem unless $\underline{y}_m$ contains $\underline{y}$. When $\underline{y}_m = \begin{bmatrix} \underline{y} \\ \tilde{\underline{y}}_m \end{bmatrix}$, $\tilde{\underline{y}}_m = \tilde{\underline{C}}_m \underline{x} + \tilde{\underline{D}}_m \underline{u}$, where without loss of generality it is assumed that $\operatorname{rank} \begin{bmatrix} \underline{C} \\ \tilde{\underline{C}}_m \end{bmatrix} = r + r_e$ where $\operatorname{rank} \tilde{\underline{C}}_m \triangleq r_e$, the following theorem is obtained:

Theorem 3

A necessary and sufficient condition that there exists a linear, time-invariant robust controller for (1) such that $\underline{e}(t) \to \underline{0}$ as $t \to \infty$ for all disturbances $\underline{\omega}_i$, $i=1,2,\ldots,\Gamma$ and all reference outputs $\underline{y}_i$, $i=1,2,\ldots,\Gamma$ described by (2) and such that the controlled system is controllable (or stabilizable) is that the following conditions all hold:

(a) $(\underline{A},\underline{B})$ be controllable (or stabilizable)

(b) $\{\underline{A},(\underline{C},\tilde{\underline{C}}_m)\}$ be observable (or detectable)

(c) $\operatorname{rank} \begin{bmatrix} \underline{A}-\lambda_j^i \underline{I} & \underline{B} \\ \underline{C} & \underline{D} \end{bmatrix} = n+r$, $j=1,2,\ldots,p^{\bar{i}}$; $i=1,2,\ldots,\Gamma$

The following robust controller will accomplish the objectives of Theorem 3:

Minimum Order Robust Controller

A robust controller which is of minimum order $(r\phi+n-r-r_e)$ for unspecified gains consists of the following now assuming that $\{\underline{A},(\underline{C},\tilde{\underline{C}}_m)\}$ is observable:

$$\underline{u} = \underline{K}(\underline{I},\underline{0},\ldots,\underline{0})\hat{\underline{\xi}} + \sum_{i=1}^{r} \underline{K}_i \underline{\eta}_i \qquad (16)$$

where

$$\dot{\underline{\eta}}_i = \underline{C}\,\underline{\eta}_i + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} e_i \ , \ i=1,2,\ldots,r \tag{17}$$

and $\underline{C}$ is a $\phi \times \phi$ matrix defined in (9) , where the $(n+r\phi)$ vector $\hat{\underline{\xi}}$ is the output of a minimum order observer with input $(\underline{e}', \tilde{\underline{y}}_m', \underline{\eta}_1', \ldots, \underline{\eta}_r')'$ which estimates the state $\underline{\xi}$ of the observable system:

$$\dot{\underline{\xi}} = \underline{A}\underline{\xi} + \underline{B}\underline{u}$$

$$\begin{bmatrix} \underline{e} \\ \tilde{\underline{y}}_m \\ \underline{\eta}_1 \\ \vdots \\ \underline{\eta}_r \end{bmatrix} = \begin{bmatrix} \underline{C} & \underline{0} & \cdots & \underline{0} \\ \tilde{\underline{C}}_m & \underline{0} & \cdots & \underline{0} \\ \underline{0} & \underline{I} & \cdots & \underline{0} \\ \vdots & \vdots & & \\ \underline{0} & \underline{0} & \cdots & \underline{I} \end{bmatrix} \underline{\xi} + \begin{bmatrix} \underline{D} \\ \tilde{\underline{D}}_m \\ \underline{0} \\ \vdots \\ \underline{0} \end{bmatrix} \underline{u} \tag{18}$$

where $\underline{A}$, $\underline{B}$ are defined in (10) and $\underline{K}$, $\underline{K}_i$, $i=1,2,\ldots,r$ are found so that $\{\underline{A} + \underline{B}(\underline{K}, \underline{K}_1, \ldots, \underline{K}_r)\}$ has specified stable dynamic behaviour, where the pair $(\underline{A}, \underline{B})$ is controllable (or stabilizable).

Remark

The modification required for the controller when $\{\underline{A}, (\underline{C}, \tilde{\underline{C}}_m)\}$ is only detectable is obvious.

Remarks

6. When $\underline{y}_m = (\underline{y}', \tilde{\underline{y}}_m', \bar{\underline{y}}_m')'$ where $\bar{\underline{y}}_m = \bar{\underline{C}}_m\underline{x} + \bar{\underline{D}}_m\underline{u} + \sum_{i=1}^{\Gamma} \bar{\underline{F}}_i^m \underline{\omega}_i$, the corresponding conditions of Theorem 3 are straightforward to obtain. It should be noted however that in this case, it may not always be advantageous to use $\bar{\underline{y}}_m$ in the controller, because one must increase the order of the observer to account for the disturbances occurring in $\bar{\underline{y}}_m$.

7. The above controller has all the properties described previously in Theorem 1 (see remarks 1,2,3).

Discussion of Theorem 3

It should be emphasized that the minimal order robust controller obtained in Theorem 3 is quite distinct from an observer [3] and can be considered to be a new type of compensator - a "*servo compensator*" . It is a unique device and cannot be thought of as being derived from an observer.

The frequency domain interpretation of Theorem 1 and 3 is interesting; it means that in order for a robust controller to exist for a system no *transmission zeros* of the system $\dot{\underline{x}} = \underline{A}\underline{x} + \underline{B}\underline{u}$, $\underline{y} = \underline{C}\underline{x} + \underline{D}\underline{u}$ can coincide with the characteristic roots λ_j^i of the differential equations describing the class of reference outputs or/and disturbance inputs (see [4] for a definition of transmission zeros).

IV. SOME NUMERICAL EXAMPLES

The following example illustrates some of the previous results: Consider the system:

$$\dot{\underline{x}} = \begin{pmatrix} -1 & 1 \\ 0 & -2 \end{pmatrix} \underline{x} + \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \underline{u} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \omega_1 + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \omega_2 + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \omega_3 \tag{19}$$

$$\underline{y} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \underline{x} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \underline{u} + \begin{pmatrix} -3 \\ 0 \end{pmatrix} \omega_1$$

where ω_1, ω_2 are step-function disturbances and ω_3 is a ramp-function disturbance. It is desired to find a minimal order robust controller assuming $\underline{x}, \underline{y}$ are available for measurement so that $y_1 \to y_{\text{ref}_1}$ $y_2 \to y_{\text{ref}_2}$ as $t \to \infty$ with specified dynamic behaviour where y_{ref_2} is a constant set point and y_{ref_1} is a sinusoidal output with frequency Ω.

$$\text{Let } \underline{A} = \begin{pmatrix} -1 & 1 \\ 0 & -2 \end{pmatrix}, \ \underline{B} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \ \underline{C} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \ \underline{D} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}; \tag{20}$$

then from Theorem 1 there is a solution if and only if:

(a) $(\underline{A}, \underline{B})$ is controllable

(b) $$\text{rank} \begin{bmatrix} \underline{A} & \underline{B} \\ \underline{C} & \underline{D} \end{bmatrix} = 4, \ \text{rank} \begin{bmatrix} \underline{A} - j\Omega \underline{I} & \underline{B} \\ \underline{C} & \underline{D} \end{bmatrix} = 4 \tag{21}$$

It is immediately verified that these conditions are satisfied. Therefore there is a solution to the problem for all $\Omega \varepsilon R^1$ and the controller is given by:

$$\underline{u} = \underline{K}_0 \underline{x} + (\underline{K}_1, \underline{K}_2) \begin{bmatrix} \underline{\xi}_1 \\ \underline{\xi}_2 \end{bmatrix} \tag{22}$$

where $$\dot{\underline{\xi}}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\Omega^2 & 0 \end{bmatrix} \underline{\xi}_1 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} (y_1 - y_{\text{ref}_1}),$$

$$\dot{\underline{\xi}}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\Omega^2 & 0 \end{bmatrix} \underline{\xi}_2 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} (y_2 - y_{\text{ref}_2}) \tag{23}$$

where $(\underline{K}_0, \underline{K}_1, \underline{K}_2)$ are found so that the controllable closed loop system has specified dynamic behaviour by using either pole assignment or optimal regulator theory (see [5]).

REFERENCES

[1] Davison, E.J., "The Feedforward and Feedback Control of a General Servomechanism Problem - Part I" Department of Elect. Engineering, Control System Report No.7305, University of Toronto, April, 1973.

[2] Davison, E.J., Wang, S.H., "Properties of Linear Multivariable Systems Subject to Arbitrary Output and State Feedback", IEEE Trans. on Automatic Control, Vol. AC-18, No.1, February 1973, pp. 24-32.

[3] Luenberger D., "Observers for Multivariable Systems", IEEE Trans. on Automatic Control, Vol. AC-11, No.2, April 1966, pp. 190-197.

[4] Wang S.H., Davison E.J., "Canonical Forms of Linear Multivariable Systems", Dept. of Electrical Engineering, University of Toronto Control System Report No.7203, April 1972, presented at 6th Asilomar Conference on Circuits & Systems, November 1972.

[5] Davison E.J., "The Output Control of Linear Time-Invariant Multivariable Systems with Unmeasurable Arbitrary Disturbances", IEEE Trans.on Automatic Control, Vol. AC-17, No.5, Oct. 1972, pp. 621-630.

DESIGN OF LOW-ORDER, DYNAMIC, FEEDBACK COMPENSATORS VIA A POLE-PLACEMENT ALGORITHM

JUDE R. ANDERS and S. H. WU
Marquette University, Milwaukee, Wisconsin

ABSTRACT

Several procedures are developed and demonstrated for the design of dynamic, feedback compensators for single input/single output, linear, constant systems with unavailable states. These procedures are based on a pole-placement-problem solution which reduces the design problem to the solution of a set of linear equations. Also, an algorithm is presented for accomplishing the design which is compatible to implementation on a digital computer.

INTRODUCTION

The problem of interest here is the design of dynamic feedback compensators for the placement of closed-loop poles. The a-priori restriction for arbitrary pole placement, usually a part of this problem formulation, [1] and [2], is not imposed. The basic idea here is to find a compensator of low order which will result in a satisfactory closed-loop pole arrangement. The system under consideration is the linear, constant, single input/single output dynamic system of order n, which has the well-known model:

$$S_1: \quad \dot{\underline{x}}(t) = A\underline{x}(t) + \underline{b}u(t) \tag{1}$$

$$y(t) = \underline{c}^T\underline{x}(t) \tag{2}$$

In addition, it is assumed that S_1 is completely controllable, and thus, represented in control canonical form. That is,

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{bmatrix}; \quad \underline{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{3}$$

$$\underline{c}^T = [c_0 \; c_1 \; \cdots \; c_{n-1}]$$

The dynamic compensator to be designed is assumed to be a system of order p, modeled as:

$$S_2: \quad \dot{\underline{z}}(t) = D\underline{z}(t) + \underline{g}y(t) \tag{4}$$

$$r(t) = \underline{h}^T\underline{z}(t) + ky(t) \tag{5}$$

where k is a scalar constant, and D, $\underline{g}$, and $\underline{h}^T$ have form similar to the corresponding elements of S_1. The resultant composite system will then be described by:

$$S_3: \quad \begin{bmatrix} \dot{\underline{x}}(t) \\ \hline \dot{\underline{z}}(t) \end{bmatrix} \begin{bmatrix} A+k\underline{b}\underline{c}^T & \underline{b}\underline{h}^T \\ \hline \underline{g}\underline{c}^T & D \end{bmatrix} \begin{bmatrix} \underline{x}(t) \\ \hline \underline{z}(t) \end{bmatrix} + \begin{bmatrix} \underline{b} \\ \hline \underline{0} \end{bmatrix} v(t) \tag{6}$$

$$y(t) = [\underline{c}^T \mid \underline{0}] \begin{bmatrix} \underline{x}(t) \\ \underline{z}(t) \end{bmatrix} \tag{7}$$

By direct manipulation, the transfer function of S_3 can be obtained in the form:

$$\frac{Y(s)}{V(s)} = \frac{(c_{n-1}s^{n-1}+\ldots+c_2s^2+c_1s+c_0)(s^p+d_{p-1}s^{p-1}+\ldots+d_1s+d_0)}{s^{n+p}+\gamma_{n+p-1}s^{n+p-1}+\ \ldots\ +\gamma_2s^2+\gamma_1s+\gamma_0} \qquad (8)$$

where the γ_i in the denominator of (8) specify the S_3 characteristic equation, i.e. the closed loop poles. Thus, the problem can be reduced to that of obtaining a relation between the 2p+1 parameters of S_2 and the n+p γ_i of (8).

PRELIMINARY RESULTS

Composite System Characteristic Equation

In [3], the form of the γ_i from (8) are revealed by developing a similarity transformation matrix, T, which transforms the S_3 system matrix into the companion matrix form, which has the γ_i as its last row. The general form of T is shown in (9). In the development of the proof that T is indeed the desired matrix, the relation (10) becomes evident. This is the same type of relation obtained by Pearson [1]. Note that (10) is a linear relation between the γ_i and the compensator parameters expressed in terms of the known S_1 system parameters. Thus, the compensator design problem is reduced to obtaining solutions to the matrix relation (10), written in short form as:

$$\underline{\gamma} = M\,\underline{\lambda} \qquad (11)$$

It can be shown [3] that if S_1 is observable and p=n-1, the square matrix M is invertible and a solution exists for arbitrary $\underline{\gamma}$. This is the pole placement result of Pearson [1]. However, the main interest here is for the case $p \leq n-1$. Such solutions are considered in the design procedures discussed below.

Optimal System Formulation

The compensator design problem has also been posed in terms of optimizing a performance index, J, of the form:

$$J = \int_0^\infty \{\, \underline{x}^T Q_1 \underline{x} + \underline{u}^T Q_2 \underline{u} + w^2 \} \, dt \qquad (12)$$

where Q_1 and Q_2 are positive semidefinite matrices, $\underline{x}$ is the S_1 state vector, $\underline{u}$ consists of the first p-derivatives of u(t) and $w=u(t)^{(p)}$ [1 & 2]. If $\underline{x}$ and $\underline{u}$ are considered the state vectors of a composite system associated with (12), system relations can be written in the form:

$$S_u: \quad \begin{aligned} \dot{\underline{x}}(t) &= A\underline{x}(t) + \underline{b}u(t) \\ \dot{\underline{u}}(t) &= B\underline{u}(t) + \underline{e}w(t) \\ y(t) &= \underline{c}^T\underline{x}(t) \end{aligned} \qquad (13)$$

The optimal linear regulator solution will yield a control law solution of the form:

$$w = -(\underline{k}_1^T \, \underline{x} + \underline{k}_2^T \, \underline{u}) \qquad (14)$$

where the $\underline{k}_i$ contain the optimal state feedback constants. Using this law, the resulting composite optimal system can be written as:

$$
\left[\begin{array}{cccccc:cccc}
h_0 & h_1 & \cdots\ h_{p-1} & 0 & \cdots & 0 & 0 & 0\ \cdots & 0 & 0 \\
0 & h_0 & \cdots\ h_{p-2} & h_{p-1} & \cdots & 0 & 0 & 0\ \cdots & 0 & 0 \\
\vdots & & & & & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & 0 & h_0 & \cdots & h_{p-1} & 0 & 0\ \cdots & 0 & 0 \\
0 & \cdots & 0\ \ 0 & h_0 & \cdots & h_{p-2} & h_{p-1} & 0\ \cdots & 0 & 0 \\
\vdots & & & & & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & h_0 & & h_1 & h_2 & h_3\cdots & 0 & 0 \\
0 & 0 & \cdots & 0 & & h_0 & h_1 & h_2\cdots & h_{p-1} & 0 \\
\hdashline
-(kc_0-a_0) & -(kc_1-a_1)\cdots & -(kc_{n-2}-a_{n-2}) & & & -(kc_{n-1}-a_{n-1}) & 1 & 0\ \cdots & 0 & 0 \\
0 & -(kc_0-a_0)\cdots & -(kc_{n-3}-a_{n-3}) & & & -(kc_{n-2}-a_{n-2}) & -(kc_{n-1}-a_{n-1}) & 1\ \cdots & 0 & 0 \\
\vdots & & & & & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots-(kc_0-a_0) & & \cdots & & -(kc_{n-p}-a_{n-p}) & -(kc_{n-p+1}-a_{n-p+1}) & \cdots-(kc_{n-1}-a_{n-1}) & & 1
\end{array}\right] \tag{9}
$$

The Transformation Matrix T

$$\begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p - a_0 \\ \gamma_{p+1} - a_1 \\ \vdots \\ \vdots \\ \gamma_{n+p-1} - a_{n-1} \end{bmatrix} = \begin{bmatrix} c_0 & 0 \cdots & 0 & 0 & a_0 & 0 \cdots & 0 \\ c_1 & c_0 \cdots & 0 & 0 & a_1 & a_0 \cdots & 0 \\ c_2 & c_1 & \vdots & \vdots & a_2 & a_1 & \vdots \\ c_3 & c_2 & 0 & 0 & a_3 & a_2 & 0 \\ \vdots & \vdots & c_0 & 0 & \vdots & \vdots & a_0 \\ c_{n-1} & c_{n-2} & c_1 & c_0 & a_{n-1} & a_{n-2} & a_1 \\ 0 & c_{n-1} & c_2 & c_1 & 1 & a_{n-1} & a_2 \\ 0 & 0 & \vdots & \vdots & 0 & 1 & \vdots \\ 0 & 0 & & & 0 & 0 & \\ \vdots & \vdots & c_{n-2} & c_{n-3} & \vdots & \vdots & a_{n-2} \\ & & c_{n-1} & c_{n-2} & & & a_{n-1} \\ 0 & 0 \cdots & 0 & c_{n-1} & 0 & 0 \cdots & 1 \end{bmatrix} \begin{bmatrix} -(h_0 + kd_0) \\ -(h_1 + kd_1) \\ -(h_2 + kd_2) \\ \vdots \\ -(h_{p-1} + kd_{p-1}) \\ -k \\ d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_{p-1} \end{bmatrix} \tag{10}$$

$$S_u^o: \quad \begin{bmatrix} \dot{\underline{x}}(t) \\ \hline \dot{\underline{u}}(t) \end{bmatrix} = \left[\begin{array}{c|c} A & \underline{b}[0] \\ \hline -\underline{e}\underline{k}_1^T & B - \underline{e}\underline{k}_2^T \end{array}\right] \begin{bmatrix} \underline{x}(t) \\ \hline \underline{u}(t) \end{bmatrix} + \begin{bmatrix} \underline{b} \\ \hline \underline{0} \end{bmatrix} v(t) \tag{15}$$

$$y(t) = [\underline{c}^T \mid \underline{0}^T] \begin{bmatrix} \underline{x}(t) \\ \hline \underline{u}(t) \end{bmatrix}$$

In [3] it is shown that the results of this optimization can be translated into a compensator design, with composite system form given by (6) and (7), by a solution of (11) where only $\underline{\gamma}$ is changed as follows:

$$\underline{\gamma} = \begin{bmatrix} \underline{k}_1 + T_3^T \underline{k}_2 \\ T_4^T \underline{k}_2 \end{bmatrix} \tag{16}$$

where T_3 is the lower left submatrix of T, (9), with $-(kc_i - a_i)$ replaced by a_i; and T_4 is the lower right submatrix of T with the same replacement of elements. In words, the solution of the optimal problem, specifies the characteristic equation of S_3 according to (16). Again, the problem is reduced to the solution of (11).

DESIGN PROCEDURES

Since the compensator design is specified by the relation (11), design procedures can be established through consideration of general solutions to (11). Two such solutions are evaluated here.

Exact Solution

Assume that M in (11) has rank $m \leq 2p+1$ where $p \leq n-1$. Then, there exists M_{11}, an $(m \times m)$ invertible kernel in M, such that by interchange of columns and rows of M with corresponding modifications of the vectors $\underline{\gamma}$ and $\underline{\lambda}$, (11) can be written as:

$$\left[\begin{array}{c} \underline{\gamma}_1 \\ \hline \underline{\gamma}_2 \end{array}\right] = \left[\begin{array}{c|c} M_{11} & M_{12} \\ \hline CM_{11} & CM_{12} \end{array}\right] \left[\begin{array}{c} \underline{\lambda}_1 \\ \hline \underline{\lambda}_2 \end{array}\right] \tag{17}$$

The matrix C represents the linear dependence of the rows not in M_{11} on those in M_{11}. It can be shown [8] that (11) has the general solution:

$$\left[\begin{array}{c} \underline{\lambda}_1 \\ \hline \underline{\lambda}_2 \end{array}\right] = \left[\begin{array}{c} M_{11}^{-1}\,\underline{\gamma}_1 - M_{11}^{-1}M_{12}\underline{\lambda}_2 \\ \hline \underline{\lambda}_2 \end{array}\right] \tag{18}$$

if and only if:

$$\underline{\gamma}_2 = C\,\underline{\gamma}_1 \tag{19}$$

This relation is well known as the consistency condition for (11).

Now, consider the results of this discussion in terms of the compensator design problem. The consistency condition specifies the set of possible characteristic equations for S_3 of given order n+p. For each characteristic equation from this set, (18) specifies compensator designs. The designer's problem is reduced to choosing a suitable element from this set. For the controllable and observable S_1 he is assured that for p=n-1, this set contains all characteristic equations of order 2n-1.

Approximate Solution

The pseudo-inverse matrix defined by Penrose [6] can be used to obtain an approximate solution to (11). That is, the pseudo-inverse, M^+, of M in (11) yields a solution:

$$\underline{\lambda}_s = M^+\,\underline{\gamma} \tag{20}$$

such that the Euclidean norm:

$$\|M\,\underline{\lambda}_s - \underline{\gamma}\| = \|\underline{\gamma}' - \underline{\gamma}\| \tag{21}$$

is minimized. Also, it can be shown [5] that:

$$M^+ = Q^*(QQ^*)^{-1}(R^*R)^{-1}R^* \tag{22}$$

where * denotes conjugate transpose and the matrices Q and R can be expressed in terms of M_{11}, CM_{11}, $M_{11}^{-1}M_{12}$ which are involved in (17) and (18) above, and the mxm identity matrix, I_m, as:

$$R = \left[\begin{array}{c} M_{11} \\ \hline CM_{11} \end{array}\right] \quad ; \quad Q = \left[\begin{array}{c|c} I_m & M_{11}^{-1}M_{12} \end{array}\right] \tag{23}$$

This solution can be used in the compensator design problem as follows. For a specified S_3 order, n+p, and specified characteristic equation ($\underline{\gamma}$), (20) yields a compensator design, such that the resultant S_3 characteristic equation ($\underline{\gamma}'$) is close to the specified one in the sense that the norm (21) is minimized. While this procedure does not guarantee the proximity of specified and resultant poles, the norm must be zero in the case of the existence of an exact solution. Also, those examples attempted revealed a tendency for the poles of S_1 to move toward the specified set. In the case that the result proves unsatisfactory, the designer may either simply specify a new $\underline{\gamma}$ or increase the order n+p and re-specify $\underline{\gamma}$. In the latter case, he is assured of suitable pole placement at p=n-1 for S_1 controllable and observable.

PIVOTAL ALGORITHM

Since the partitioning shown in (17) is not a simple matter for large n+p, an algorithm is developed which facilitates the design procedures discussed above. This algorithm, a modification of the symmetric method due to Talacko [4] for solution of linear programming problems, is actually an algorithm for the solution of a set of linear equations. It proceeds in an orderly manner from a standard array in the form:

$$\begin{array}{c|c|c} & \lambda^T & \\ \hline 1 & & \\ 2 & & \\ \vdots & M & \underline{\gamma} \\ \vdots & & \\ n+p & & \\ \hline \end{array} \qquad (24)$$

to a final array in the form:

$$\begin{array}{c|c:c|c} & & \underline{\lambda}_2^T & \\ \hline \underline{\lambda}_1 & M_{11}^{-1} & M_{11}^{-1}M_{12} & M_{11}^{-1}\underline{\gamma}_1 \\ \hdashline & -C & 0 & \underline{\gamma}_2 - C\underline{\gamma}_1 \\ \hline \end{array} \qquad (25)$$

where the partitioning is obviously marked by those λ_i transferred in the procedure. A comparison of this final array and the relations (17), (18), and (23) demonstrates that both the exact and approximate solutions discussed above can be obtained via simple steps from information presented on the array. The procedure of the algorithm is presented below in the example to add to the clarity of its presentation.

EXAMPLE PROBLEM

Suppose the given system S_1 is described by:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & -3 & -3 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u \qquad (26)$$

$$y = [1 \quad 0 \quad 0 \quad 0]\ \underline{x} \qquad (27)$$

This system has poles at s=-2,-1, and ±j and characteristic equation:

$$S_1(s) = s^4 + 3s^3 + 3s^2 + 3s + 2 \qquad (28)$$

Also, suppose compensators of order p=1 are to be considered. For this case, (11) is written as:

$$\begin{bmatrix} \gamma_0 \\ \gamma_1 - 2 \\ \gamma_2 - 3 \\ \gamma_3 - 3 \\ \gamma_4 - 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -(h_0+kd_0) \\ -k \\ d_0 \end{bmatrix} \tag{29}$$

Then, the procedure of the pivotal algorithm is applied as follows:

1.) Form the standard array (array 0 below).

2.) Choose a pivot element (shown circled) from the non-zero elements of M and form a new array by performing the following operations:
   a. Interchange the elements on the top of the pivot column and left of the pivot row.
   b. Replace the pivot by its reciprocal.
   c. Divide each element of the pivot row by the pivot.
   d. Divide each element of the pivot column by the pivot and multiply by -1.
   e. All remaining elements are replaced by the resultant of: e-q•r/p
   where: e is the element being replace.
   q is the element at the intersection of the pivot row and e's column.
   r is the element at the intersection of the pivot column and e's row.
   p is the pivot.

3.) Pivot selection from the center of each new array and array formation according to #2 proceeds in sequence until no further transfer of λ_i's from the top to the left is possible. That is, once the maximum number of λ_i's have been transferred, the array formation is complete.

4.) Reorder the final array placing transferred elements in upper left portion and all elements in numerical order according to the indices at top and left. The result is an array of the form (25).

$$\begin{array}{c|ccc|l} 0 & \lambda_1 & \lambda_2 & \lambda_3 & \\ \hline 1 & 1 & 0 & 2 & \gamma_0 \\ 2 & 0 & 1 & 3 & \gamma_1-2 \\ 3 & 0 & 0 & \textcircled{3} & \gamma_2-3 \\ 4 & 0 & 0 & 3 & \gamma_3-3 \\ 5 & 0 & 0 & 1 & \gamma_4-3 \\ \hline \end{array} \Longrightarrow \begin{array}{c|ccc|l} 1 & \lambda_1 & \lambda_2 & 3 & \\ \hline 1 & \textcircled{1} & 0 & -2/3 & \gamma_0 -2/3\, \gamma_2 - 2 \\ 2 & 0 & \textcircled{1} & -1 & \gamma_1 - \gamma_2 + 1 \\ \lambda_3 & 0 & 0 & 1/3 & \gamma_2/3 - 1 \\ 4 & 0 & 0 & -1 & \gamma_3 - \gamma_2 \\ 5 & 0 & 0 & -1/3 & \gamma_4 - \gamma_2/3 - 2 \\ \hline \end{array} \Longrightarrow$$

$$\begin{array}{c|ccc|l} 3 & 1 & 2 & 3 & \\ \hline \lambda_1 & 1 & 0 & -2/3 & \gamma_0 - 2/3\, \gamma_2 + 2 \\ \lambda_2 & 0 & 1 & -1 & \gamma_1 - \gamma_2 + 1 \\ \lambda_3 & 0 & 0 & 1/3 & 1/3\, \gamma_2 - 1 \\ \hdashline 4 & 0 & 0 & -1 & \gamma_3 - \gamma_2 \\ 5 & 0 & 0 & -1/3 & \gamma_4 - 1/3\, \gamma_2 - 2 \\ \hline \end{array}$$

Clearly, the unity pivots in array #1 accomplish only a transfer of λ_i-elements. Array #3 is the final array in ordered form. Thus, the following can be stated.

TABLE I.

Examples of Modeling a Characteristic Equation

Parameters	Binomial Form		Butterworth Form		ITAE Form	
p=1	A	B	A	B	A	B
γ_0	1.000	1.000	1.000	1.000	1.000	1.000
γ_1	5.000	5.000	3.240	3.240	3.400	3.400
γ_2	10.000	9.948	5.250	5.169	5.500	5.100
γ_3	10.000	9.948	5.250	5.169	5.000	5.100
γ_4	5.000	5.316	3.240	3.723	2.800	3.700
p=2						
γ_0	1.000	1.000	1.000	1.000	1.000	1.000
γ_1	6.000	6.000	3.860	3.860	3.950	3.950
γ_2	15.000	15.000	7.460	7.460	7.450	7.450
γ_3	20.000	20.019	9.130	9.177	8.600	8.682
γ_4	15.000	14.933	7.460	7.325	6.600	6.360
γ_5	6.000	6.130	3.860	4.133	3.250	3.733

A: Specified Coefficients, B: Resultant Coefficients

TABLE II.

Pseudo-Inverse Solution With Specification Of The Optimal Characteristic Equation

Compensator Order	Optimal Gains $k_i, i=1 \ldots n+p$	Specified Characteristic Eq. $\gamma_i, i=0 \ldots n+p-1$	Resultant Characteristic Eq. $\gamma_i, i=0 \ldots n+p-1$
p=1	0.196	3.112	3.112
	2.124	8.498	8.498
	2.908	10.282	9.216
	0.980	8.354	9.216
	1.458	4.458	5.072
p=2	-1.216	2.236	2.236
	-1.550	7.850	7.850
	-0.175	13.336	13.336
	0.159	14.670	14.667
	1.726	11.059	11.069
	2.111	5.111	5.090

The matrix M has rank 3, i.e., full column rank and M^+ is the generalized left inverse of M. The consistency condition here requires that:

$$\gamma_3 = \gamma_2 \tag{30}$$

$$\gamma_4 = 1/3\,\gamma_2 + 2 \tag{31}$$

Thus the compensator design specified by (18) is:

$$\begin{bmatrix} h_0+kd_0 \\ -k \\ d_0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -2/3 \\ 0 & 1 & -1 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1-2 \\ \gamma_2-3 \end{bmatrix} \tag{32}$$

and will give a resultant S_3 characteristic equation:

$$S_3(s) = s^5+(\gamma_2/3+2)s^4+\gamma_2 s^3+\gamma_2 s^2+\gamma_1 s+\gamma_0 \tag{33}$$

Compensators for the system (26) were computed for p=1 and 2. Designs using the approximate solution (20) where $\underline{\gamma}$ was specified by the binomial, Butterworth, and ITAE standard forms [7], are shown in Table I, comparing the specified and resultant parameters. Similarly, the optimal formulation (16) was used to specify $\underline{\gamma}$ for solution (20). The results are summarized on Table II. The performance index used here is given by:

$$J = \int_0^\infty \left\{ \sum_{i=1}^{4} x_i^2 + \sum_{i=1}^{p} u_i^2 + \dot{u}_p^2 \right\} dt \tag{34}$$

CONCLUSION

Several design procedures have been developed and demonstrated for low order feedback compensators. Furthermore an algorithm for computing the design has been presented. This algorithm is suitable for use on a digital computer. Extension of the procedure to discrete-time systems is obvious. However, work in determining the pole trajectories as a function of the parameters involved in the approximate solution could lead to interesting results.

REFERENCES

[1] J.B. Pearson, "Compensator Design for Dynamic Optimization", Int. J. Contr., Vol. 9, pp. 473-482, 1969.

[2] J.B. Pearson and C.Y. Ding, "Compensator Design for Multivariable Linear Systems", IEEE Trans. Auto. Contr., Vol. AC-14, pp. 130-134, April, 1969.

[3] J.R. Anders, "A Low-Order Compensation Technique", Master's Thesis, Marquette U., Dept. of Electrical Engineering, Milwaukee, Wis., 1971.

[4] J. Talacko, Class Notes, Marquette U., Dept. of Mathematics, Milwaukee, Wis., 1971.

[5] R. K. Manherz, "Applications of the Generalized Matrix Inverse in Network Theory", Ph.D. Thesis, Technological Institute, Northwestern U., Evanston, Ill., 1968.

[6] R. Penrose, "A Generalized Inverse for Matrices", Proc. Cambridge Philos. Soc., Vol. 51, pp. 406-413, 1955.

[7] J.J. D'Azzo and C.H. Houpis, Feedback Control System Analysis and Synthesis, McGraw-Hill, New York, pp. 589-592, 1966.
[8] S.R. Searle, Matrix Algebra for the Biological Sciences, John Wiley & Sons, Inc., New York, pp. 136-143, 1966.

A SEQUENTIAL ALGORITHM FOR CALCULATION OF OPTIMAL OUTPUT-CONSTRAINED FEEDBACK REGULATORS

S.P.BINGULAC
Control and Computation Laboratory, Institute "B. Kidric" - Vinca
Belgrade, Yugoslavia, presently on leave at the University of Brasilia,
Department of Physics, 70.000 Brasilia-DF, Brasil
N.M.ĆUK
Serbian Electric Power Company
Belgrade, Yugoslavia
M.S.ĆALOVIĆ
Serbian Electric Power Company
Belgrade, Yugoslavia

ABSTRACT

The paper considers the problem of numerical calculation of optimal output-constrained feedback regulators for linear plants. It has been experienced that existing computational procedures for solving this type of problems do not assure satisfactory results. This is particularly true when the dimension of an output vector is considerably lower than the order of the system.

A new computational procedure is formulated introducing appropriate modifications to the original problem. It assures the convergence even when the existing computational algorithms fail. The main characteristic of this procedure is that it yields the solution in several steps. The number of steps depends on the difference between the order of the system and the dimension of the output vector. The first step gives the solution of the full state feedback regulator, while the last one provides the solution of a specified control problem for an optimal regulator constrained to use only the given output variables.

The paper presents a derivation of a computational algorithm and reports some numerical results illustrating feasibility of its application.

INTRODUCTION

In their recent paper [1], Levine and Athans have developed an optimal regulator constrained to use only the specified output variables for a control feedback. This regulator is much more appropriate for a practical realization than regulators with a full state feedback. However, the design of an output constrained feedback regulator involves a more complicated computational algorithm, than that for the usual Riccati equation employed in calculating the optimal gains for the full state feedback regulators. The first computational algorithms for the solution of this optimal control problem were suggested in [1,2] and proved on low-order examples. Examining these algorithms on the examples of high-order systems, it has been found that they cannot provide satisfactory results, particularly in the cases when the number of outputs is much smaller than the order of the system. It has been concluded that the main reason for this failure is an inability to find an adequate initial guess for the iterative procedure when solving the associated system of algebraic matrix equations.

A new computational algorithm proposed in this paper avoids the difficulties inherent in the previously developed algorithms. It consists of solving a sequence of control problems, starting from a solution of a full state feedback regulator. The last control problem in the sequence represents the specified output constrained feedback regulator. In this sequence each preceding control problem is used as an initial guess for the solution of the next one. Thus, in order to find an adequate initial guess for the solution of a specified regulator problem, the proposed computational procedure requires the solution of several auxiliary control problems.

In the derivation of this computational procedure, the following two modifications of the previously proposed algorithms are introduced:

a. an appropriate transformation of related matrix expressions that facilitates a solution of a specified control problem;

b. a systematic method for a selection of initial guesses when searching for the optimum feedback gains.

These modifications allow an easy application of the standard Newton-Raphson algorithm in the compact Kronecker product form.

To illustrate the applicability of the proposed procedure, the paper refers to some numerical results obtained in a study of the 19-th order system where the number of output variables was taken to be 9 or 5.

STATEMENT OF THE PROBLEM

As shown by Levine and Athans [1], for a linear time-invariant, completely controllable and completely observable system

$$S_{pm}\left\{\begin{array}{l} \dot{x}(t) = Ax(t) + Bu(t);\ x(o) = x_0, \\ y(t) = Cx(t), \end{array}\right. \tag{1}$$

where $x \in R^n$, $u \in R^m$, $y \in R^p$ are state, control and output vectors, respectively, while $\{A,B,C\}$ (Rank $C=p$) are constant system matrices of compatible dimensions, an optimum output-constrained feedback regulator is of the form

$$u(t) = -Fy(t). \tag{2}$$

The constant mxp dimensional feedback gain matrix F in (2) is obtained by minimizing a quadratic performance index

$$J\ (x_o,F) = \frac{1}{2}\int_0^\infty [(x^T(t)Qx(t) + u^T(t)Ru(t)]\ dt, \tag{3}$$

where the constant weighting matrices Q and R have the following properties:

$$Q = Q^T \geqslant 0; \quad R = R^T > 0.$$

As shown in [1], when this optimal control problem has a solution, the feedback gain is given by

$$F = R^{-1}B^TKLC^T(CLC^T)^{-1}, \tag{4}$$

where nxn dimensional matrices K and L are positive definite solutions of algebraic matrix equations

$$KA_c + A_c^T K + Q + C^T F^T R F C = 0 \quad ; \tag{5}$$

$$LA_c^T + A_c L + I = 0, \tag{6}$$

while $A_c = A - BFC$ is assumed to be a stable matrix.

Equations (4)-(6) are obtained by a minimization of expected value $E\left\{J_{min}(x_o,F)\right\}$ over all x_o, assuming that x_o is a random variable uniformly distributed over the n-dimensional unit sphere. This insures an elimination of the dependence of the feedback gain F on the initial state x_o. Then the expected value of the minimal performance index (3) reduces to

$$J(F) = E\left\{J_{min}(x_o,F)\right\} = \frac{1}{2n}\, tr[K] \, . \tag{7}$$

Numerical procedures for a solution of a coupled nonlinear system of equations (4)-(6) were also suggested in [1,2]. Since the convergence is not guaranteed for none of these algorithms, they failed in some examples of high-order systems where $p \ll n$. This fact was the primary motivation for a research of a new computational algorithm.

TRANSFORMATION OF THE SYSTEM (4)-(6)

To obtain a new computational algorithm, the system (4)-(6) should be transformed into a more convenient form. By introducing substitutions

$$H = C^T(CLC^T)^{-1}C \quad \text{and} \quad G = BR^{-1}B^T, \tag{8}$$

the equations (5), (6) and (8) may be expressed by the following system of three nxn nonlinear matrix equations

$$E^1(K,L,H) = KA + A^TK + Q - KGKLH - HLKGK + HLKGKLH = 0; \tag{9}$$

$$E^2(K,L,H) = AL + LA^T + I - LHLKG - GKLHL = 0; \tag{10}$$

$$E^3(K,L,H) = H - C^T(CLC^T)^{-1}C = 0. \tag{11}$$

Without loss of generality, it is assumed that the matrix C in (1) is of the form*

$$C = \left[I_p \;\vdots\; 0 \right] \tag{12}$$

---

*If the matrix C in the triple $\{A,B,C\}$ defining the system (1) is not given by (12), it can be converted into the desired form by the nonsingular transformation [3]

$$T^{-1} = \begin{bmatrix} C \\ \hdashline C_a \end{bmatrix},$$

where C_a is arbitrary (n-p)xn matrix, assuring the regularity of T. Then the same system (1) is defined by a new triple $\{\bar{A},\bar{B},\bar{C}\}$, where $\bar{A}=T^{-1}AT$; $\bar{B}=T^{-1}B$; $\bar{C}=CT=\left[I_p \;\vdots\; 0\right]$, while corresponding weighting matrices $\bar{Q}$ and $\bar{R}$ are $\bar{Q}=Q^TQT$ and $\bar{R}=R$.

Accordingly, the matrix H given by (8) becomes

$$H = \left[\begin{array}{c|c} H_a & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} L_a^{-1} & 0 \\ \hline 0 & 0 \end{array}\right], \qquad (13)$$

where pxp dimensional matrix L_a is defined by the following partition of matrix L

$$L = \left[\begin{array}{c|c} L_a & L_b^T \\ \hline L_b & L_c \end{array}\right], \qquad (14)$$

so that

$$H_a L_a = I_p. \qquad (15)$$

Thus, the nxn matrix E^3 in (11) may be reduced to a pxp dimensional matrix expression

$$E^4 (L_a, H_a) = H_a L_a - I_p = 0. \qquad (16)$$

Taking into account a symmetric property of matrices E^1, E^2, E^4, K,L and H_a, the system of nonlinear matrix equations (9), (10) and (16) can be considered as a system of k nonlinear scalar equations

$$e(z) = 0; \quad z \in R^k; \quad e \in R^k, \qquad (17)$$

where unknown vector z contains all independent elements (upper triangle) of symmetric matrices K,L and H_a. Their total number is

$$k = 2\,\frac{n(n+1)}{2} + \frac{p(p+1)}{2}. \qquad (18)$$

As shown in the sequel, an explicit derivation of the vector expression (17) for the transformed system (9), (10), (16) is not necessary. It can be performed by using the Kronecker product techniques [5].

Assuming an appropriate initial guess for the unknown vector z exists, system (17) may be solved applying any of the existing reliable computational procedures [4].

THE PROBLEM OF INITIAL GUESS

In a previously developed algorithm [2], it was suggested to use the solution of the full feedback regulator problem as an initial guess in the iterative procedure when solving (4)-(6). It has been experienced that convergence problems in the solution of (4)-(6) particularly appear in high order systems when $p \ll n$. Then, it seems more natural to arrive to an adequate initial guess by solving a sequence of n-p auxiliary optimal control problems. The sequence starts with a solution of the standard Riccati equation, to obtain optimum gains for the full state feedback regulator. Then, gradually decreasing the dimension of output vector y from n to p, and using the solution of the preceding problem as an initial guess for the next one, the solution of the specified output constrained regulator can be obtained in (n-p+1)-st step of the sequence.

Each of the control problems in the suggested sequence may be defined as follows:

For an n-th order system S_{jm}, with m inputs and j outputs, $j \in [n,p]$

$$S_{jm} \begin{cases} \dot{x}^j(t) = Ax^j(t) + Bu^j(t) \; ; \\ y^j(t) = C^j x^j(t); \quad C^j = \left[I_j \mid 0 \right] , \end{cases} \tag{19}$$

find an optimum output feedback F^j that defines a linear control law

$$u^j(t) = -F^j y^j(t) , \tag{20}$$

minimizing a quadratic performance index of the form (3).

From (9), (10), (16) and (19) it follows that mxj dimensional feedback gain matrix is given by

$$F^j = R^{-1} B^T K^j L^j \begin{bmatrix} H_a^j \\ \hline 0 \end{bmatrix} , \tag{21}$$

where matrices K^j, L^j and H_a^j satisfy the following system of nonlinear matrix equations:

$$E^{1j} = K^j A + A^T K^j + Q - K^j G K^j L^j H^j - H^j L^j K^j G K^j + H^j L^j K^j G K^j L^j H^j = 0; \tag{22}$$

$$E^{2j} = AL^j + L^j A^T + I_n - GK^j L^j H^j L^j - L^j H^j L^j K^j G = 0; \tag{23}$$

$$E^{4j} = H_a^j L_a^j - I_j = 0. \tag{24}$$

Matrices H_a^j and L_a^j are defined by the relations (13) and (14) for each particular $j \in [n,p]$.

The first optimal control problem corresponding to j = n involves the solution of Riccati and Lyapunov algebraic matrix equations. These solutions may be obtained applying the existing reliable computational procedures [6,7] . The values of matrices K^n and L^n obtained in the first step are used as initial guesses for the second one characterized by j = n-1, and so on. The last (n-p+1)-st step gives the solution of the specified output-constrained feedback regulator, since, according to (12)

$$C^{j=p} = C = \left[I_p \mid 0 \right] \tag{25}$$

Since matrices K^j and L^j are obtained by solving the same system of nonlinear matrix equations (22)-(24), where the only difference is in the dimension of the jxj matrix H_a^j, containing all nonzero elements of H^j, it may be concluded that the matrix couple $\{K^{j+1}, L^{j+1}\}$ will be relatively close to $\{K^j, L^j\}$. This experimentally proved fact justifies the use of solutions from the preceding step as an initial guess for the next one.

NEWTON-RAPHSON APPROACH TO THE SOLUTION OF THE SYSTEM (22), (23), (24)

For each $j \in [n,p]$, the feedback gain F^j is obtained by solving the corresponding system (22)-(24). To simplify the notation, the explicit dependence on the index j will be omitted.

Expanding equations (22)-(24) into Taylor series about $\{K,L,H_a\} = \{K_i,L_i,H_{ai}\}$ for

$$\{K,L,H_a\} = \{K_{i+1}, L_{i+1},(H_a)_{i+1}\} = \{K_i + \Delta K_i, L_i + \Delta L_i, H_{ai} + \Delta H_{ai}\}$$

and neglecting the second and higher order terms, yields

$$\begin{aligned}&\Delta K_i(A-GK_iL_iH_i)+(A-GK_iL_iH_i)^T\Delta K_i+(H_iL_i-I)K_iG\Delta K_iL_iH_i+H_iL_i\Delta K_iGK_i(H_iL_i-I)^T+\\&+(H_iL_i-I)K_iGK_i\Delta L_iH_i+H_i\Delta L_iK_iGK_i(H_iL_i-I)^T+(H_iL_i-I)K_iGK_iL_i\Delta H_i+\\&+\Delta H_iL_iK_iGK_i(H_iL_i-I)^T=-E_i^1\,;\end{aligned}\tag{26}$$

$$\begin{aligned}&-L_iH_iL_i\Delta K_iG - G\Delta K_iL_iH_iL_i + \Delta L_i(A-GK_iL_iH_i)^T + (A-GK_iL_iH_i)\Delta L_i -\\&-L_iH_i\Delta L_iK_iG - GK_i\Delta L_iH_iL_i - L_i\Delta H_iL_iK_iG - GK_iL_i\Delta H_iL_i = -E_i^2;\end{aligned}\tag{27}$$

$$(\Delta L_a)_i(H_a)_i + (L_a)_i(\Delta H_a)_i = -E_i^4,\tag{28}$$

$$\text{where } \Delta H_i = \left[\begin{array}{c|c}(\Delta H_a)_i & 0\\ \hline 0 & 0\end{array}\right];\quad \Delta L_i = \left[\begin{array}{c|c}(\Delta L_a)_i & (\Delta L_b)_i^T\\ \hline (\Delta L_b)_i & (\Delta L_c)_i\end{array}\right].\tag{29}$$

Using (28), the matrix $(\Delta H_a)_i$ may be eliminated in (26) and (27), so that the total number of unkonwns reduces to $N = 2\frac{n(n+1)}{2}$. Assuming the matrix $(H_a)_i$ is calculated in each iteration from (13) and in view of (16) $(H_a)_i = (L_a)_i^{-1}$, the increment $(\Delta H_a)_i$ can be expressed as

$$(\Delta H_a)_i = -(H_a)_i(\Delta L_a)_i(H_a)_i.\tag{30}$$

According to (29), this gives

$$\Delta H_i = -H_i\Delta L_i H_i\,.\tag{31}$$

By substituting (31) into (26) and (27), after the rearrangements, the following system of two nxn coupled linear matrix equations is obtained

$$(\Delta KA_c+A_c^T\Delta K+W\Delta KV^T+V\Delta KW^T-WKV^T\Delta LH-H\Delta LVKW^T)_i=-E_i^1\tag{32}$$

$$(-LHL\Delta KG-G\Delta KLHL+\Delta LA_c^T+A_c\Delta L+V^T\Delta LW+W^T\Delta LV)_i=-E_i^2\tag{33}$$

$$\begin{aligned}\text{where}\quad &V_i = (HL - I)_i;\\ &W_i = (HL - I)_iK_iG = V_iK_iG\,;\\ &A_{ci} = A - GK_i.\end{aligned}\tag{34}$$

The iterative scheme based on the Newton-Raphson techniques

$$\{K_{i+1}, L_{i+1}\} = \{K_i,L_i\} + \{\Delta K_i, \Delta L_i\}$$

possesses the known convergence properties [4] and provides a rapid convergence in the neighborhood of the solution satisfying (26)-(28).

Procedures for Solving the System of Equations (32)-(33)

The straightforward procedure for solving (32)-(33) consists in the application of Kronecker product techniques [5], to expand matrix equations (32)-(33) into the system of $N = 2\frac{n(n+1)}{2}$ scalar equations written in the vector form (17).

$$U^i \Delta z_i = e_i, \tag{35}$$

where

$$U^i = \left[\begin{array}{c|c} U^i_{11} & U^i_{12} \\ \hline U^i_{21} & U^i_{22} \end{array}\right]; \quad \Delta z_i = \left[\begin{array}{c} \Delta k_i \\ \hline \Delta l_i \end{array}\right]; \quad e_i = \left[\begin{array}{c} e^1_i \\ \hline e^2_i \end{array}\right].$$

$\Delta k_i, \Delta l_i, e^1_i$ and e^2_i are $\frac{N}{2}$ dimensional vectors containing independent elements (upper triangle) of symmetric matrices $\Delta K_i, \Delta L_i, E^1_i$ and E^2_i respectively, i.e.

$$\Delta k^T_i = [\Delta k_{11}, \Delta k_{12}, \ldots, \Delta k_{1n}, \Delta k_{22}, \Delta k_{23}, \ldots, \Delta k_{2n}, \Delta k_{33}, \Delta k_{34}, \ldots, \Delta k_{nn}]_i;$$

$$\Delta l^T_i = [\Delta l_{11}, \Delta l_{12}, \ldots, \Delta l_{1n}, \Delta l_{22}, \Delta l_{23}, \ldots, \Delta l_{2n}, \Delta l_{33}, \Delta l_{34}, \ldots, \Delta l_{nn}]_i;$$

$$\cdots\cdots \qquad \cdots\cdots \qquad \cdots\cdots \qquad .$$

Then, the solution of (35) requires the inversion of the NxN matrix U^i at each iteration.

The $\frac{N}{2} \times \frac{N}{2}$ dimensional matrices U^i_{11}, U^i_{12}, U^i_{21} and U^i_{22} in (35) are given by relations

$$\begin{aligned} U^i_{11} &\triangleq (A^T_c \overset{\cdot}{\underset{\cdot}{\times}} I_n + I_n \overset{\cdot}{\underset{\cdot}{\times}} A^T_c + W \overset{\cdot}{\underset{\cdot}{\times}} V + V \overset{\cdot}{\underset{\cdot}{\times}} W)_i \; ; \\ U^i_{12} &\triangleq (-H \overset{\cdot}{\underset{\cdot}{\times}} VKW^T - WKV^T \overset{\cdot}{\underset{\cdot}{\times}} H)_i \; ; \\ U^i_{21} &\triangleq (-G \overset{\cdot}{\underset{\cdot}{\times}} LHL - LHL \overset{\cdot}{\underset{\cdot}{\times}} G)_i \; ; \\ U^i_{22} &\triangleq (A_c \overset{\cdot}{\underset{\cdot}{\times}} I_n + I_n \overset{\cdot}{\underset{\cdot}{\times}} A_c + W^T \overset{\cdot}{\underset{\cdot}{\times}} V^T + V^T \overset{\cdot}{\underset{\cdot}{\times}} W^T)_i = (U^i_{11})^T, \end{aligned} \tag{36}$$

where matrices A_{ci}, V_i and W_i are given by (34).

The symbol "$\overset{\cdot}{\underset{\cdot}{x}}$" designates the symmetric Kronecker matrix product [5].

In the solution of (32)-(33), the main problem is to derive a computational procedure which possibly does not require the inversion of large matrices. Some preliminary results describing a method which handles only nxn dimensional matrices are reported in [8].

DESCRIPTION OF COMPUTATIONAL ALGORITHMS

The sequential algorithm is organized at three levels, as illustrated in Figure 1.

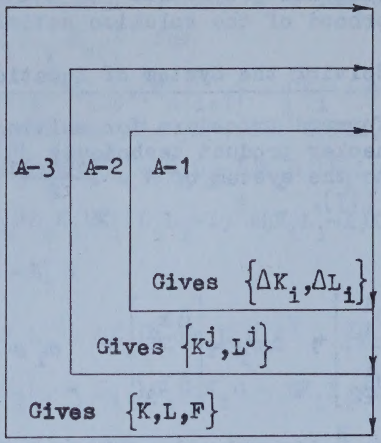


Figure 1. The Organization of Computational Algorithm

The first level deals with a solution of equations (32)-(33) giving $\{\Delta K_i, \Delta L_i\}$ that satisfy the equation (35) (Algorithm A-1).

At the second level, optimization of the system S_{jm}(19) is performed. It gives $\{K^j, L^j\}$ that satisfy the equations (22)-(24) (Algorithm A-2).

The third level comprises the sequential algorithm for the optimization of the system S_{pm}(1). It gives the solutions $\{K, L, F\}$ that satisfy the equations (9)-(11) as well as (4)-(6) (Algorithm A-3).

Iterative Algorithm A-1 for the Solution of Equations (32)-(33)

Step 1. Set $r = 0$, $\Delta K_i^r = 0$, $\Delta L_i^r = 0$, $e_k \ll 1$; $e_l \ll 1$.

Step 2. Set $r = r+1$.

Step 3. Having ΔL_i^{r-1} find ΔK_i^r by solving the following linear matrix equation:

$$(\Delta K^r A_c + A_c^T \Delta K^r + W \Delta K^r V^T + V \Delta K^r W^T + S_1^r)_i = 0,$$

where
$$(S_1^r)_i = (E^1 - WKV^T \Delta L^{r-1} H - H \Delta L^{r-1} VKW^T)_i \tag{37}$$

while matrices A_{ci}, V_i and W_i are defined by (34).

Step 4. Having ΔK_i^r, find ΔL_i^r by solving the following linear matrix equation:

$$(\Delta L^r A_c^T + A_c \Delta L^r + W^T \Delta L^r V + V^T \Delta L^r W + S_2^r)_i = 0$$

where
$$(S_2^r)_i = (E^2 - LHL \Delta K^r G - G \Delta K^r LHL)_i. \tag{38}$$

Step 5. If $\| \Delta K_i^r - \Delta K_i^{r-1} \| \leqslant e_k$ and

$\left\| \Delta L_i^r - \Delta L_i^{r-1} \right\| \leqslant e_1$ stop; else, go to step 2.

Linear matrix equation to be solved in steps 3 and 4 of the above algorithm is called "extended" Lyapunov matrix equation [9] because of its similarity to the classical linear Lyapunov matrix equation [7] . The convergent algorithm for solving the extended Lyapunov matrix equations is described in [8] .

Iterative Algorithm A-2 for the Optimization of System S_{jm} (19)

Step a. Set i=1; $\left\{K_i^j, L_i^j\right\} = \left\{K^{j+1}, L^{j+1}\right\}$; $e_1 \ll 1$; $e_2 \ll 1$.

Step b. Having $\left\{K_i^j, L_i^j\right\}$ find $\left\{\Delta K_i^j, \Delta L_i^j\right\}$ solving either system (35) or applying the algorithm A-1.

Step c. If $\left\|\Delta K_i^j\right\| \ll e_1$ and $\left\|\Delta L_i^j\right\| \ll e_2$, go to step d; else, set

$$\left\{K_{i+1}^j, L_{i+1}^j\right\} = \left\{K_i^j + \Delta K_i^j, L_i^j + \Delta L_i^j\right\}; \quad i = i+1$$

and go to step b.

Step d. Set $\left\{K^j, L^j\right\} = \left\{K_i^j, L_i^j\right\}$ and stop.

The couple $\left\{K^j, L^j\right\}$ obtained in step d satisfies the equations (22)-(24).

Sequential Algorithm A-3 for the Optimization of System S_{pm} (1)

Step I. Define the realization $\left\{A,B,C\right\}$ of S_{pm} (1) and corresponding weighting matrices $\left\{Q,R\right\}$ satisfying (12), and set $j = n$.

Step II. Calculate K^j solving the standard Riccati matrix equation $K^jA + A^TK^j - K^jGK^j + Q = 0$.

Step III. Having K^j, calculate L^j from linear Lyapunov matrix equation

$$L^jA_{cj}^T + A_{cj}L^j + I = 0,$$

where

$$A_{cj} = A - GK^j.$$

Step IV. Set $j = j - 1$.

Step V. Using $\left\{K^{j+1}, L^{j+1}\right\}$ as the initial guess, apply the algorithm A-2 for the optimization of the system S_{jm}. The algorithm gives $\left\{K^j, L^j\right\}$.

Step VI. If $j = p$ set $\left\{K,L\right\} = \left\{K^p, L^p\right\}$ and stop; else go to step IV.

APPLICATION

This computational algorithm has been applied to the problem of automatic generation control of an interconnected power system. The system model was of the 19-th order, with two inputs and 9 or 5 outputs. The application of the algorithm proposed in [2] failed to converge, while the sequential algorithm gave the solution. The full description of the mathematical model and the interpretation of the obtained results are given in [10].

CONCLUSIONS

This paper presents a new computational procedure for a numerical solution of the problem of optimum linear regulators with incomplete state feedback. The proposed sequential algorithm has proven to be more successfull than previously suggested algorithms particularly in the case of high order systems, when the number of output variables is small compared to the system order. The applicability of this algorithm is illustrated on an example of a large scale power system where the previously proposed algorithms failed.

REFERENCES

1. W.S.Levine, M. Athans, "On the Determination of the Optimal Constant Output Gains for Linear Systems", IEEE Trans. on AC, Vol. AC-15, No.1, Feb. 1970, pp 44-48.

2. W.S. Levine, T.L. Johnson, M. Athans, "Optimum Limited State Variable Feedback Controllers for Linear Systems", IEEE Trans. on AC, Vol. AC-16, Dec. 1971, pp. 785-793.

3. S. Bingulac and M. Djorović, "On the Generation of Sensitivity and Pseudosensitivity Functions", Proceedings, IFAC Symposium on Sensitivity, Adaptivity and Optimality, Ischia, Italy, June 18 to 21, 1973.

4. E.Polak, "Computational Methods in Optimization", Academic press, New York, N.Y., 1971.

5. G.Singh, "Optimization and Decoupling of Large Dynamic Systems with Applications to Power Systems", Ph.D. thesis, University of Illinois, Urbana, Ill., 1971.

6. J.E.Potter, "Matrix Quadratic Solutions", SIAM J. Appl. Math., Vol. 14, No. 3, May 1966, pp. 496-501.

7. R.A.Smith, "Matrix Equation XA+BX=C", SIAM J. Appl. Math, Vol. 16, No. 1, Jan. 1968, pp. 198-201.

8. S.P.Bingulac, N.M.Ćuk, "Computational Algorithm for Solving Extended Lyapunov Equations", Publication of the Institute "B. Kidrič" - Vinča, Beograd, 1973.

9. S.P.Bingulac, N.M.Ćuk, "On the Solution of Linear Matrix Equation GKH+H'KG=-W", Proceedings, Int.Conf. on Systems and Controls, Aug. 30 - Sept. 1, 1973, Coimbatore, India.

10. M.S.Ćalović, S.P.Bingulac, N.M.Ćuk, "An Output Feedback Proportional-Plus-Integral Regulator for Automatic Generation Control", IEEE PES Summer Meeting, Vancouver, B.C.,Can.,July 1973, Paper No. C73-489-2.

THE OPTIMAL LINEAR REGULATOR WITH APERIODIC SAMPLING TIMES

by

Robert A. Schlueter
Michigan State University

ABSTRACT

The optimal linear regulator with aperiodic sampling times is formulated, the existence of an optimal open-loop control law is proved, and an efficient search algorithm is developed. The problem is reformulated to obtain a closed loop control law. The search algorithm is then used to compute the open and closed loop control for a second-order system. The resulting sampling criteria are shown to be optimal and provide a basis for evaluating periodic and arbitrary aperiodic sampling criteria.

I. Introduction

Periodic sampling criteria have often been used in sampled-data control systems in order to simplify the design and analysis. Aperiodic sampling criteria have become quite practical with the introduction of computers in both design and control. Therefore, numerous aperiodic sampling criteria have been studied[1-8] in an effort to improve system performance and sampling efficiency relative to a periodic sampling criterion.

Optimal sampled-data control problems have been formulated[9, 10] in an effort to obtain sampling criteria that provide better performance than any periodic or arbitrary aperiodic sampling criterion. Necessary conditions were derived in both papers, but have never been used to obtain an efficient computational algorithm for the optimal solution. A sequential unconstrained minimization technique has been used[11] with some success in the special case where the continuous-time problem can be transformed to an equivalent discrete-time one.

II. Open Loop Control Law[12]

Consider a time invariant system

$$\dot{\underline{x}}(t) = \underline{A}\underline{x}(t) + \underline{B}\underline{u}(t) \tag{1}$$

$$\underline{y}(t) = \underline{C}\underline{x}(t) \tag{2}$$

$$\underline{x}(t_o) = \underline{\xi} \tag{3}$$

with quadratic cost functional

$$J = \frac{1}{2}\underline{x}'(t_f)\,\underline{Q}\underline{x}(t_f) + \frac{1}{2}\int_{t_o}^{t_f} [\underline{x}'(t)\underline{\hat{Q}}\underline{x}(t) + \underline{u}'(t)\underline{R}\underline{u}(t)]\, dt \tag{4}$$

where the dimension of $\underline{x}$, $\underline{y}$, and $\underline{u}$ are n, m, and r, respectively; t_o is the known initial time and $t_f \in (t_o, \infty)$ is the free terminal time. Moreover, a sampling constraint restricts the control to be a piecewise constant vector function whose elements change value only at the sampling times

$$\{t_{i+1}\}_{i=0}^{N-1}; \quad 0 < a_i \leq t_{i+1} - t_i = T_i \leq b_i \tag{5}$$

$$\underline{u}(t) = \underline{u}(t_i) = \underline{u}_i \qquad \text{for all } t \in [t_i, t_{i+1}]; \tag{6}$$

for i = 0, 1, . . ., N-1. Determine the sequences

$$\{\underline{u}_i\}_{i=0}^{N-1} \qquad\qquad \{T_i\}_{i=0}^{N-1}$$

that minimize (4) subject to (1), (2), (3), (5) and (6).

The sampled-data problem can be transformed into an equivalent discrete-time one[12] by integrating (1) and (4) over each sampling interval $T_i = t_{i+1} - t_i$.

$$\underline{x}_{i+1} = \underline{\Phi}_i \underline{x}_i + \underline{D}_i \underline{u}_i \tag{7}$$

$$J = \frac{1}{2} \underline{x}'_N \underline{Q} \underline{x}_N + \frac{1}{2} \sum_{i=0}^{N-1} (\underline{x}'_i \underline{Q}_i \underline{x}_i + 2\underline{x}'_i \underline{M}_i \underline{u}_i + \underline{u}'_i \underline{R}_i \underline{u}_i) \tag{8}$$

where $\underline{x}_i = \underline{x}(t_i)$ and

$$\underline{\Phi}_i = \underline{\Phi}(T_i) \tag{9a}$$

$$\underline{D}_i = \underline{D}(T_i) = \int_0^{T_i} \underline{\Phi}(t) \underline{B} dt \tag{9b}$$

$$\underline{Q}_i = \underline{Q}(T_i) = \int_0^{T_i} \underline{\Phi}'(t) \underline{\hat{Q}} \underline{\Phi}(t) dt \tag{9c}$$

$$\underline{M}_i = \underline{M}(T_i) = \int_0^{T_i} \underline{\Phi}'(t) \underline{\hat{Q}} \underline{D}(t) dt \tag{9d}$$

$$\underline{R}_i = \underline{R}(T_i) = \int_0^{T_i} [\underline{R} + \underline{D}'(t) \underline{\hat{Q}} \underline{D}(t)] \, dt \tag{9e}$$

The matrices $\underline{M}_i$, $\underline{R}_i$ and $\underline{Q}_i$ are in general time varying even though $\underline{Q}$ and $\underline{R}$ are constant because the sampling intervals are not equal. The matrix $\underline{\Phi}_i$ is nonsingular because it is a fundamental matrix. Moreover, it is easily shown that $\underline{Q}_i[\underline{R}_i]$ is a positive semidefinite (definite) symmetric matrix because $\underline{Q}[\underline{R}]$ is a positive semidefinite (definite) symmetric matrix.

The discrete-time problem becomes:

Given the sampled-data system (7) with specified initial condition (3), determine the control and sampling period sequences

$$\{\underline{u}_i\}_{i=0}^{N-1} \quad \text{and} \quad \underline{T}' = (T_0, T_1, \ldots, T_{N-1})$$

that minimize the cost function (8) subject to the sampling period constraints (5).

The following theorem establishes both the existence of an optimal solution and the structure for the computational algorithm which follows.

Theorem 1

An optimal open loop control law exists for the optimal linear regulator with aperiodic sampling times.

Proof: A feasible solution exists for each $\underline{T} \in [\underline{a}, \underline{b}]$ since the state equations (7) are linearly independent for all $\underline{T} \in [\underline{a}, \underline{b}]$. Therefore, it follows from Theorems 1 and 3 [13, pp. 137 and 133, respectively] that there exist unique control and trajectory sequences

$$\{\underline{u}_i(\underline{T})\}_{i=0}^{N-1} \qquad \{\underline{x}_{i+1}(T)\}_{i=0}^{N-1} \tag{10}$$

for each $\underline{T} \in [\underline{a}, \underline{b}]$ since matrices $\underline{R}_i$ are positive definite for i=0, 1, ..., N-1 and all $\underline{T} \in [\underline{a}, \underline{b}]$. The necessary conditions [12] can be solved to obtain the control law

$$\underline{u}_i(\underline{T}) = - \underline{G}_i(\underline{T}) \underline{x}_i(\underline{T}) \tag{11}$$

where the r x n dimensional feedback gain matrix satisfies

$$\underline{G}_i(\underline{T}) = \underline{R}_i^{-1} \underline{M}'_i + [\underline{R}_i + \underline{D}'_i \underline{K}_{i+1} \underline{D}_i]^{-1} \underline{D}'_i \underline{K}_{i+1} \underline{\Theta}_i \tag{12}$$

The matrices $\underline{K}_i(\underline{T})$ satisfy the matrix Riccati equation

$$\underline{K}_i = \underline{\Gamma}_i + \underline{\Theta}'_i \left[\underline{K}_{i+1} - \underline{K}_{i+1}\underline{D}_i[\underline{R}_i + \underline{D}'_i\underline{K}_{i+1}\underline{D}_i]^{-1} \underline{D}'_i\underline{K}_{i+1}\right] \underline{\Theta}_i \quad (13)$$

for i = N-1, N-2, . . . , 1, 0 with boundary condition

$$\underline{K}_N = \underline{Q} \quad (14)$$

where $\underline{\Theta}_i = \underline{\Phi}_i - \underline{D}_i\underline{R}_i^{-1}\underline{M}'_i$ and $\underline{\Gamma}_i = \underline{Q}_i - \underline{M}_i\underline{R}_i^{-1}\underline{M}'_i$. The associated optimal cost for each $\underline{T} \in [\underline{a}, \underline{b}]$ can be written as a derived cost function

$$J_o(\underline{T}) = \{ J_{\underline{T}} = \tfrac{1}{2}\underline{x}_o\,\underline{K}_o(\underline{T})\,\underline{x}_o : \underline{a} \le \underline{T} \le \underline{b}\} \quad (15)$$

since the control and trajectory (10) and thus $J_{\underline{T}}$ are unique for each $\underline{T} \in [\underline{a}, \underline{b}]$.

Since $\underline{\Phi}_i$, $\underline{D}_i$, $\underline{Q}_i$, $\underline{M}_i$, and $\underline{R}_i$ are continuous matrix functions of $\underline{T}$ for i = 0, 1, . . . , N-1, matrices $\underline{K}_i(\underline{T})$ and $\underline{G}_i(\underline{T})$ are continuous matrix functions of $\underline{T}$ and the cost function (15) is a continuous function of $\underline{T}$. Since the cost function $J_o(\underline{T})$ is continuous on a closed bounded set $[\underline{a}, \underline{b}]$, there exists an optimal solution $\underline{T}^*$ for this derived problem. Thus, there exists a solution

$$\{\underline{u}_i(\underline{T}^*)\}_{i=0}^{N-1}, \quad \{\underline{x}_i(\underline{T}^*)\}_{i=1}^{N}, \quad \text{and } \underline{T}^*$$

for the optimal linear regulator with aperiodic sampling times. This control law is open loop since the derived cost function and thus $\underline{T}^*$ depends on the initial state $\underline{x}_o$.

This theorem shows that the solution $\underline{T}^*$ to the derived optimization problem (minimize $J_o(\underline{T})$ over set $[\underline{a}, \underline{b}]$) can be used to determine the optimal control and trajectory from the state equation (7) and the control law (11) after matrices $\underline{K}_i(\underline{T}^*)$ and $\underline{G}_i(\underline{T}^*)$ have been computed for i= N, N-1,..., 1, 0 from equations (12) - (14).

A sequential unconstrained minimization technique [SUMT] is now used to determine $\underline{T}^*$. A penalty function p(T) is adjoined to form an augmented cost function

$$L(\underline{T}, v) = J_o(\underline{T}) + v\,p(\underline{T}) \quad (16)$$

$$p(\underline{T}) = \sum_{i=0}^{N-1} [\min(0, T_i - b_i)]\,(T_i - b_i) + [\min(0, a_i - T_i)]\,(a_i - T_i) \quad (17)$$

which incorporates the sampling period constraint (5).

Since the penalty for violating the sampling period constraint is proportional to v(v>0), the minimization of $L(\underline{T}, v_k)$ for a monotonically increasing sequence $\{v_k\}$ results in a sequence $\{\underline{T}_k\}$ that can be proved [14] to converge to the optimal sampling period sequence $\underline{T}^*$. The exterior SUMT algorithm developed by Fiacco and McCormick [14] was used to compute $\underline{T}^*$.

The search algorithm developed in this paper has several advantages over the SUMT algorithm applied directly to the equivalent discrete-time problem [11]. First, the convergence rate is increased because the search dimension is reduced from N(n+r+1) to N and the Nn equality constraints (7) are eliminated. Second, the optimal sampling period sequence can be determined by a simple gradient search algorithm in the special case [10] where the sampling period constraint (5) can be ignored ($\underline{a} \cong \underline{0}$, $\underline{b} \cong \infty$). Third, the structure of the algorithm has shown that an optimal closed loop control law can be found by reformulating the problem so that the resulting derived cost function does not depend on the initial state $\underline{x}_o$.

There are several approaches in the literature [15, 16] that could be used to obtain a closed loop control. The following problem, where the initial conditions are randomly distributed, illustrates the theoretical development and computational algorithm which could be used to obtain a particular closed loop control law.

III. Closed Loop Control Law

Consider the linear time-invariant system

$$\dot{\underline{x}}(t) = \underline{A}\,\underline{x}(t) + \underline{B}\,\underline{u}(t) \tag{18}$$

$$\underline{y}(t) = \underline{C}\,\underline{x}(t) \tag{19}$$

with randomly distributed initial conditions

$$E\{\underline{x}_o\} = \underline{\xi}_o \tag{20}$$

$$E\{(\underline{x}_o - \underline{\xi}_o)(\underline{x}_o - \underline{\xi}_o)'\} = \underline{W} \tag{21}$$

and deterministic cost functional

$$J = \frac{1}{2} E_{\underline{x}_o} \{\underline{x}'(t_f)\underline{Q}\,\underline{x}(t_f) + \int_{t_o}^{t_f} [\,\underline{x}'(t)\,\hat{\underline{Q}}\,\underline{x}(t) + \underline{u}'(t)\,\underline{R}\,\underline{u}(t)]\,dt\} \tag{22}$$

where the dimensions of $\underline{x}$, $\underline{y}$, and $\underline{u}$ are n, m, and r, respectively; t_o is the known initial time and $t_f \in (t_o, \infty)$ is the free terminal time. The sampling constraint restricts the control to be a piecewise constant vector function whose elements change value only at the sampling times.

$$\{t_i\}_{i=1}^{N} \; ; \quad 0 < a_i \leq t_{i+1} - t_i = T_i \leq b_i \tag{23}$$

$$\underline{u}(t) = \underline{u}(t_i) = \underline{u}_i \qquad \text{for } t \in [t_i, t_{i+1}) \tag{24}$$

for i=0, 1, 2, ... , N-1. Determine the sequences

$$\{\underline{u}_i\}_{i=0}^{N-1} \qquad \{T_i\}_{i=0}^{N-1}$$

that minimize (22) subject to (18), (19), (20), (21), (23), and (24).

Integrating the vector differential equation (18) and the cost function (22) over each sampling interval $[t_i, t_{i+1})$ separately and invoking the control constraint (24), the continuous-time problem is transformed to the following discrete-time one:

Given the linear system and cost functional

$$\underline{x}_{i+1} = \underline{\Phi}_i \underline{x}_i + \underline{D}_i \underline{u}_i \tag{25}$$

$$J = \frac{1}{2} E_{\underline{x}_o} \{\underline{x}'_N \underline{Q}\,\underline{x}_N + \sum_{i=0}^{N-1} (\underline{x}'_i \underline{Q}_i \underline{x}_i + 2\underline{x}'_i \underline{M}_i \underline{u}_i + \underline{u}'_i \underline{R}_i \underline{u}_i)\} \tag{26}$$

with a randomly distributed initial state satisfying (20) and (21), determine

$$\{\underline{u}_i\}_{i=0}^{N-1} \qquad \text{and } \underline{T}' = (T_o, T_1, \ldots, T_{N-1})$$

that minimize this cost functional subject to the sampling period constraint (23).

The following theorem establishes both the existence of a closed loop control law and the structure for the computational algorithm.

Theorem 2

An optimal closed loop control law exists for the optimal linear regulator with aperiodic sampling times and randomly distributed initial state.

Proof: A unique closed loop control law (11) with associated cost (15) exists [12, 13] for each $\underline{T} \in [\underline{a}, \underline{b}]$ and any initial state since matrices $\underline{R}_i$ are nonsingular for i = 0, 1, 2, ..., N-1 and each $\underline{T} \in [\underline{a}, \underline{b}]$. Letting $\{\underline{u}_i(\underline{x}_o)\}_{i=0}^{N-1}$ represent any closed loop control law, the following exchange of operators is valid:

$$J_1(\underline{T}) = \min_{\{\underline{u}_i(\underline{x}_o)\}_{i=0}^{N-1}} \{J\} = \frac{1}{2} E_{\underline{x}_o} \left\{ \min_{\{\underline{u}_i(\underline{x}_o)\}_{i=0}^{N-1}} \{\underline{x}'_N \underline{Q}\underline{x}_N + \sum_{i=0}^{N-1} (\underline{x}'_i \underline{Q}_i \underline{x}_i + 2\underline{x}_i \underline{M}_i \underline{u}_i + \underline{u}'_i \underline{R}_i \underline{u}_i)\} \right\} \tag{27}$$

and thus the cost function becomes

$$J_1(\underline{T}) = \frac{1}{2} E_{\underline{x}_o} \{\underline{x}'_o \underline{K}_o(\underline{T})\underline{x}_o\} .$$

Using the identity

$$\underline{x}'_o \underline{K}_o(\underline{T})\underline{x}_o = Tr\{\underline{K}_o(\underline{T})\underline{x}_o \underline{x}'_o\} , \tag{28}$$

were $Tr\{\cdot\}$ represents the trace operator, and substituting (20), (21), the derived cost function is

$$J_1(\underline{T}) = \frac{1}{2}\left(Tr\{\underline{K}_o(\underline{T})\,\underline{W}\} + \underline{\xi}'_o\, \underline{K}_o \underline{\xi}_o \right) \tag{29}$$

The derived optimization problem (minimize $J_1(\underline{T})$ over the set $[\underline{a}, \underline{b}]$) has a solution since $\underline{K}_o(\underline{T})$ and thus $J_1(\underline{T})$ is continuous over the closed bounded set $[\underline{a}, \underline{b}]$. The control law (11) is closed loop because the derived cost function does not depend explicitly on a particular initial state.

This theorem has shown that the solution $\underline{T}^*$ to the derived optimization problem can be used to compute the optimal control law (11) after matrices $\underline{G}_i(\underline{T}^*)$ are calculated for $i=0, 1, \ldots, N-1$ using equations (12) - (14).

The SUMT algorithm is used to solve the derived optimization problem. An augmented cost function

$$L(\underline{T}, v) = J_1(\underline{T}) + v\, p(\underline{T}) \tag{30}$$

with the penalty function (17) is minimized repeatedly for a sequence $\{v_k\}$. The resulting solution sequence $\{\underline{T}_k\}$ can be shown to converge to the optimal sampling period sequence $\underline{T}^*$.

The parameters N, $\underline{a}$, and $\underline{b}$ must be chosen based on performance, communication system, and computational requirements. The number of sampling intervals must be selected based on a trade-off between system performance and computational requirements. The minimum sampling period constraint must be chosen based on communication system capacity and online computational requirements.

IV. Results

The search algorithm developed for the open and closed loop control laws has been programmed in FORTRAN IV on a CDC 6500 computer. The program uses a Fletcher-Powell conjugate gradient technique to compute the sampling interval sequence $\underline{T}^*$ that minimizes the derived cost function $J_o(\underline{T})[J_1(\underline{T})]$ for the open [closed] loop control law.

Consider the linear system

$$\dot{\underline{x}}(t) = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \underline{x}(t) + \begin{bmatrix} 0 \\ 15 \end{bmatrix} u(t) \tag{31}$$

with cost functional

$$J = \frac{1}{2}\underline{x}'(t_f) \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \underline{x}(t_f) + \frac{1}{2}\int_{t_o}^{t_f} \left(\underline{x}(t) \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \underline{x}(t) + u^2(t) \right) dt \tag{32}$$

with specified initial state $\underline{x}(t_o) = [\,5, 1\,]'$ and free terminal time. (33)

The control is piecewise constant and changes value only at the sampling times $\{t_i\}_{i=1}^{3}$ such that

$$u(t) = u(t_i) = u_i \qquad t \in [t_i, t_{i+1}) \text{ for } i = 0, 1, 2. \tag{34}$$

The Fletcher-Powell algorithm computes the optimal sampling interval sequence

$$\underline{T}^{*\prime} = (0.078, 1.572. 3.422) \tag{35}$$

which minimizes the derived cost function (15). The algorithm converges in seven iterations. The program terminates when the norm of the gradient becomes sufficiently small (< 0.002) and takes only nine seconds of CDC 6500 central processing time.

The optimal open loop control, computed using equations (7) and (11), becomes

$$u^*(t) = \begin{cases} -4.97 & : 0.000 \le t - t_o < 0.078 \\ -0.10 & : 0.078 \le t - t_o < 1.650 \\ -0.00007 & : 1.650 \le t - t_o < 5.057 \end{cases} \tag{36}$$

with associated cost $J^* = 26.968$. (37)

Now consider this example with a periodic sampling criterion over the same control interval where the sampling period is

$$T(N) = (T_o^* + T_1^* + T_2^*) \,|\, N \tag{38}$$

Letting N = 3, the periodic sampled-data control law becomes

$$u^*(t) = \begin{array}{lll} -(0.050, \ 0.050)\, \underline{x}_o & = -.30 & : 0.000 \le t - t_o < 1.686 \\ -(0.049, \ 0.047)\, \underline{x}_1 & = .07 & : 1.686 \le t - t_o < 3.372 \\ -(0.041, \ 0.042)\, \underline{x}_2 & = -.0008 & : 3.372 \le t - t_o < 5.057 \end{array} \tag{39}$$

with associated cost $J^* = 42.675$. The trajectories for both the periodic and optimal aperiodic sampling criteria are shown in Figures 1 and 2. A comparison of these trajectories and the values of the performance index indicates the significant performance advantage of the optimal aperiodic sampling criterion for the same number of control changes.

The value of the performance index for the optimal periodic sampled-data regulator decreases monotonically and approaches the value for the continuous-time problem as N increases. More than fifty optimal periodic control changes (N=50) are required to approximate the performance of the optimal aperiodic control with three control changes since

$$J\left(T(50)\right) = 27.000 \tag{40}$$

where $T(50) = .1014$. Moreover, this sampling period approaches the value of the smallest sampling interval for the optimal aperiodic criterion. This optimal aperiodic sampled-data control can be used to either increase the performance level for the same number of control changes or decrease the number of control changes for the same level of performance.

The example is now reformulated to obtain the closed loop control law.

Consider the linear system (31) with cost functional

$$J = \frac{1}{2} E_{\underline{x}_o} \left\{ \underline{x}'(t_f) \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \underline{x}(t_f) + \int_{t_o}^{t_f} \left(\underline{x}'(t) \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \underline{x}(t) + u^2(t) \right) dt \right\} \tag{41}$$

with randomly distributed initial state

$$E\{\underline{x}_o\} = \underline{0} \quad ; \quad E\{\underline{x}_o\,\underline{x}_o'\} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{42}$$

where t_o is specified and t_f is free.

The control is assumed to be piecewise constant and satisfies (34).

The Fletcher-Powell algorithm computes the optimal sampling interval sequence

$$\underline{T}^{*\prime} = (0.078,\ 1.52,\ 3.44) \tag{43}$$

that minimizes the derived cost function (29). The algorithm converges in six iterations and takes six seconds of computer time.

The optimal control law (11) becomes

$$u^*(t) = \begin{matrix} -(.827,\ .829)\underline{x}_o : & 0.0 \le t - t_o < .078 \\ -(.049,\ .049)\underline{x}_1 : & .078 \le t - t_o < 1.60 \\ -(.024,\ .024)\underline{x}_2 : & 1.60 \le t - t_o < 5.05 \end{matrix} \tag{44}$$

with associated cost $J^* = 1.109$.

The cost functionals (15) and (29) weigh the n eigenvalues, $\lambda_i(\underline{T})$ of $\underline{K}_o(\underline{T})$, differently and therefore the optimal sampling interval sequences (35) and (43) are different. Since the eigenvalues $\lambda_i(\underline{T})$ depend on the parameters of the system (1) and cost functional (4), the variation in the optimal sampling interval sequence for variations in the orientation (open-loop) or dispersion (closed-loop) of the initial state may be small as in this example.

V. Conclusions

The optimal linear regulator with aperiodic sampling times provides the best sampled-data (sample and hold) approximation to the continuous-time control law. Since the sampling intervals are chosen to optimize the cost functional for a given system with given sampling interval constraints, the resulting sampling criterion can be considered optimal for the optimal sampled-data control problem with N sampling intervals. Therefore, having selected the number of sampling intervals and thus the value of the performance index for the optimal sampling criterion, the performance of a periodic[12] or state-dependent[17] sampling criterion can be evaluated by comparing either the relative value of the performance index for the same number of control changes, or the number of control changes required to obtain the same performance index value. Both of these measures of sampling criterion performance depend on the number of control changes specified for the optimal sampling criterion. This concept of an optimal sampling criterion can form the basis for evaluating sampling criteria rather than the arbitrary performance indices used in the past[7,18,19].

The optimal linear regulator with aperiodic sampling times can improve performance and sampling efficiency relative to any periodic or arbitrary aperiodic sampling criterion. This performance advantage is particularly important for programmed control applications since both the number of control changes N and the number of memory locations (Nr) can be drastically reduced. Although the number of control changes can be reduced, the number of memory locations (Nn r) for the closed-loop control law may be excessive for some applications. The additional computation required to compute the control is in general not excessive and thus the optimal linear regulator with constrained sampling times can provide improved performance and sampling efficiency for a wide range of applications.

REFERENCES

[1] R. A. Volz and L. F. Kazda, "Design of a Digital Controller for a Tracking Telescope," IEEE Trans. on Automatic Control, Vol. AC-12, pp. 359-367, April 1967.

[2] R. C. Dorf, M. C. Farren, and C. A. Phillips, "Adaptive Sampling for Sampled-Data Control Systems," IRE Trans. on Automatic Control, Vol. AC-7, pp. 38-47, January 1962.

[3] R. Tomovic and G. A. Bekey, "Adaptive Sampling Based on Amplitude Sensitivity," IEEE Trans. on Automatic Control, Vol. AC-11, pp. 282-284, April 1966.

[4] D. Tabak, "Optimization of Nuclear Reactor Fuel Recycle Via Linear and Quadratic Programming," IEEE Trans. on Nuclear Science, Vol. NS-15, pp. 60-64, 1968.

[5] D. Ciscator and L. Mariani, "On Increasing Sampling Efficiency by Adaptive Sampling," IEEE Trans. on Automatic Control, Vol. AC-12, p. 318, June 1967.

[6] A. W. Bennett and A. P. Sage, "Discrete System Sensitivity and Variable Increment Sampling," Proceedings Joint Automatic Control Conference, pp. 603-612, June 1967.

[7] K. E. Tait, "Evaluation of Signal Dependent Sampling," International Journal on Control, Vol. 4, No. 4, pp. 201-239, 1966

[8] K. E. Tait, "Further Sampling Interval Criteria and their Evaluation in Discrete-Continuous Feedback Control Systems," International Journal on Control, Vol. 4, No. 4, pp. 297-324, 1966.

[9] R. A. Volz, "Optimal Control of Discrete Systems with Constrained Sample Times," Proceedings Joint Automatic Control Conference, pp. 16-21, June 1967.

[10] B. W. Jorden and E. Polak, "Optimal Control of Aperiodic Discrete-Time Systems," J. SIAM Control, Ser. A, Vol. 2, No. 3, pp. 332-345, 1965.

[11] B. C. Kuo and D. Tabak, "A Nonlinear Programming Approach to Optimal Control of Discrete Systems with Constrained Sampling Times," Proceedings Joint Automatic Control Conference, pp. 918-922, 1968.

[12] P. Dorato and A. H. Levis, "Optimal Linear Regulators: The Discrete-Time Case," IEEE Trans. on Automatic Control, Vol. AC-16, pp. 613-620, December 1971.

[13] M. D. Canon, C. D. Cullum, Jr., and E. Polak, Theory of Optimal Control and Mathematical Programming, McGraw-Hill, New York, 196

[14] A. V. Fiacco and G. P. McCormick, Nonlinear Programming: Sequential Unconstrained Minimization Techniques, Wiley, New York, 1968.

[15] W. S. Levine, T. L. Johnson, and M. Athans, "Optimal Limited State Variable Feedback Controllers for Linear Systems," IEEE Trans. on Automatic Control, Vol. AC-16, pp. 785-792, December, 1971.

[16] K. P. Dabke, "Suboptimal Linear Regulators with Incomplete State Feedback," IEEE Trans. on Automatic Control, Vol. AC-15, pp. 120-122, February 1970.

[17] R. A. Schlueter and A. H. Levis, "The Optimal Linear Regulator with State-Dependent Sampling," Proceedings Sixth Annual Princeton Conference on Information Sciences and Systems, pp. 165-169, 1972.

[18] M. J. Smith, Jr., "An Evaluation of Adaptive Sampling," IEEE Trans. on Automatic Control, Vol. AC-16, pp. 282-284, June 1971.

[19] A. H. Levis, M. Athans, and R. A. Schlueter, "On the Behavior of Optimal Linear Sampled-Data Regulators," International Journal on Control, Vol. 13, pp. 343-361, February 1971.

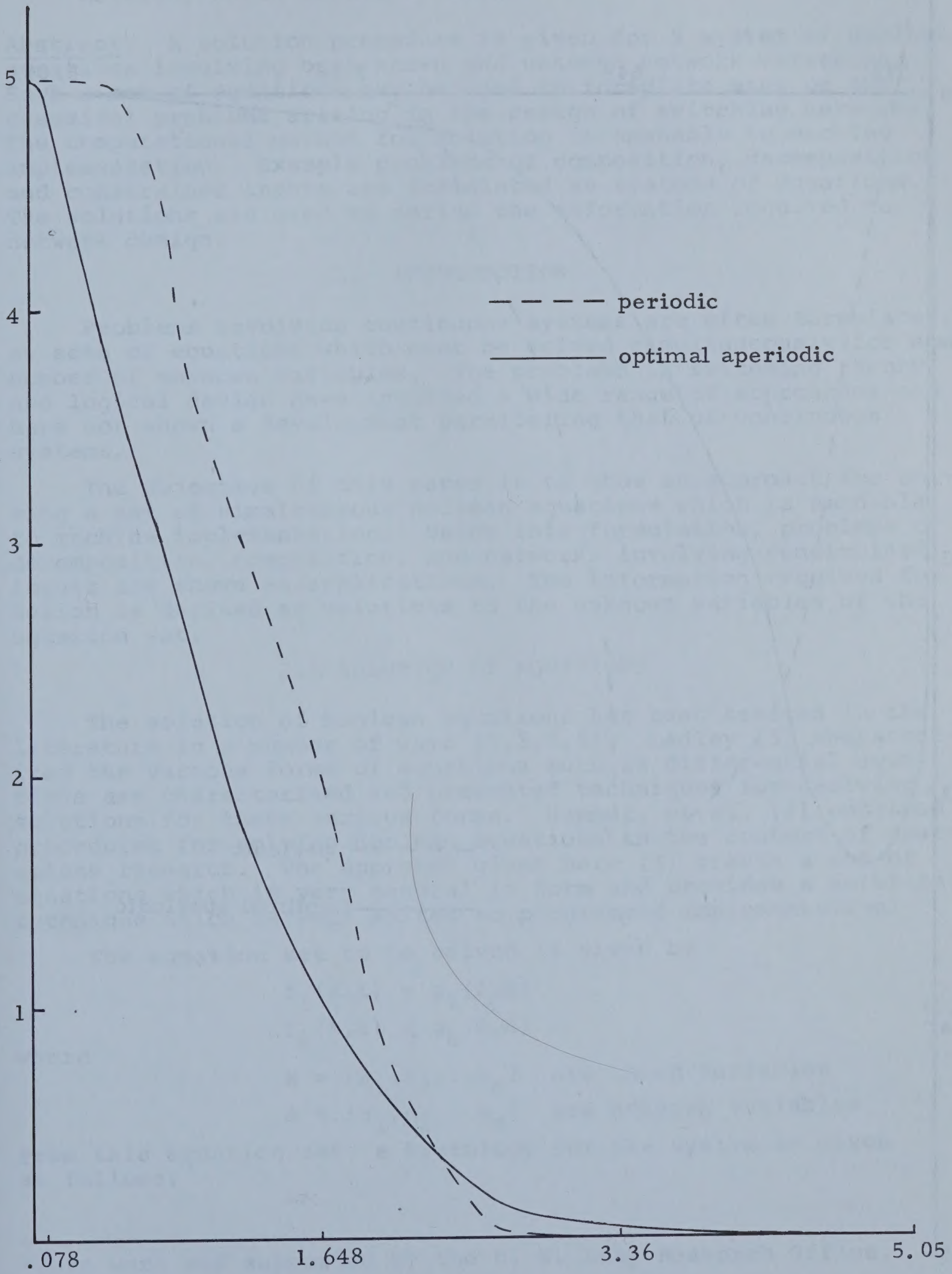


Fig. 1. Trajectory $x_1(t)$ for the optimal aperiodic and periodic sampling criteria.

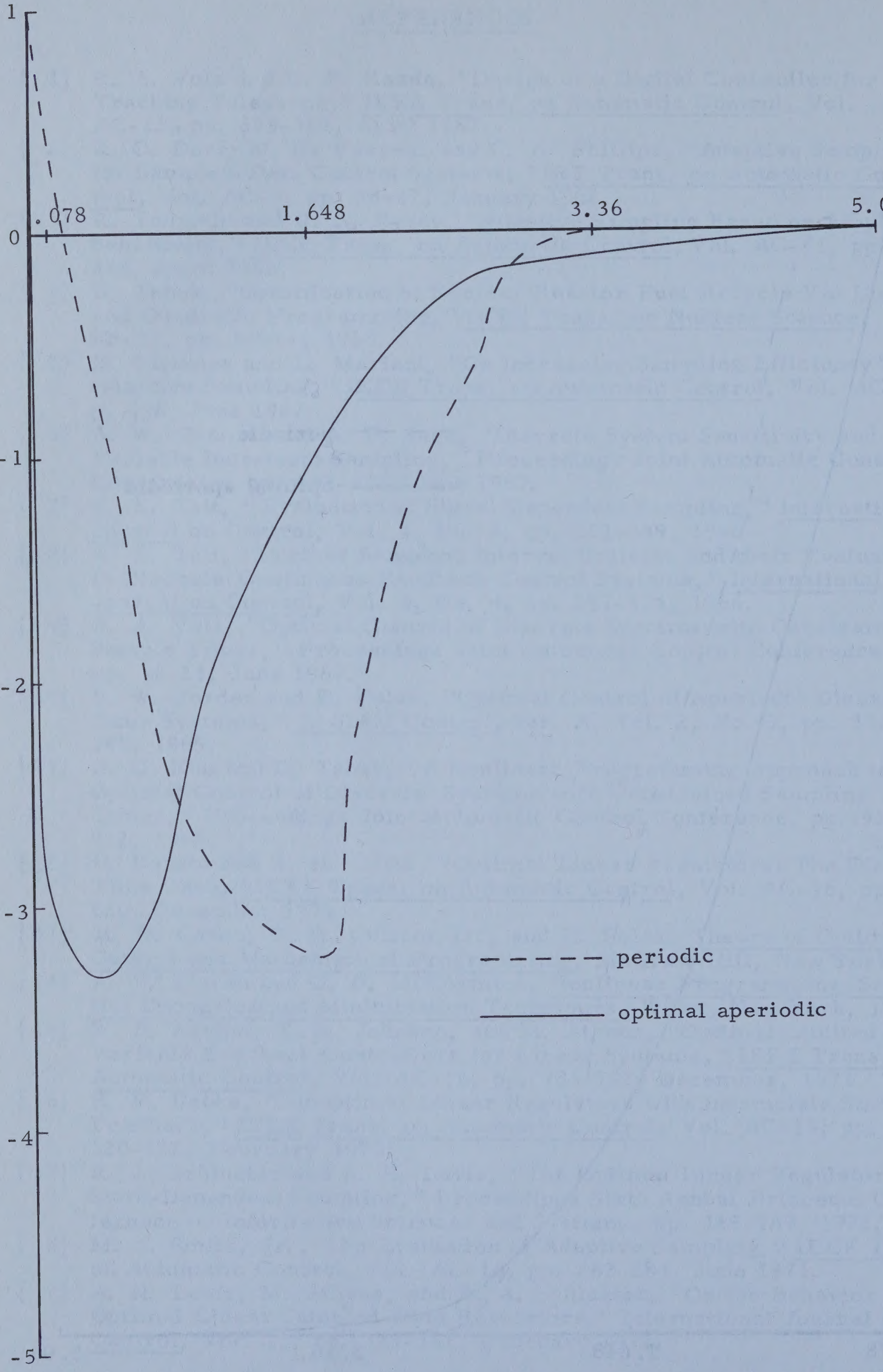


Fig. 2. Trajectory $x_2(t)$ for the optimal aperiodic and periodic sampling criteria.

THE SOLUTION OF BOOLEAN EQUATIONS IN DESIGNING SWITCHING NETWORKS

J. W. GAULT*
Electrical Engineering Department
North Carolina State University
Raleigh, North Carolina 27607

Abstract: A solution procedure is given for a system of Boolean equations involving both known and unknown network variables. Such a set of equations may be used to formulate many of the classical problems arising in the design of switching networks. The computational method for solution is amenable to machine implementation. Example problems of composition, decomposition and constrained inputs are formulated as systems of equations. The solutions are used to derive the information required for network design.

1. INTRODUCTION

Problems involving continuous systems are often formulated as sets of equations which must be solved simultaneously for some number of unknown variables. The problems in switching theory and logical design have involved a wide range of approaches and have not shown a development paralleling that of continuous systems.

The objective of this paper is to show an approach for solving a set of simultaneous Boolean equations which is amenable to machine implementation. Using this formulation, problems of decomposition, composition, and networks involving constrained inputs are shown as applications. The information required for design is derived as solutions to the unknown variables of the equation set.

2. SOLUTION OF EQUATIONS

The solution of Boolean equations has been treated in the literature in a number of ways [1,3,4,5]. Ledley [5] characterized the various forms of equations much as differential equations are characterized and presented techniques for deriving solutions for these various forms. Hammer, et.al. [4] outlined procedures for solving Boolean equations in the context of operations research. The approach given here [3] treats a set of equations which is very general in form and provides a solution technique which is well suited to programmed implementation.

The equation set to be solved is given by

$$\begin{array}{ccc} f_1(A,X) & = & g_1(A,X) \\ \vdots & \vdots & \vdots \\ f_k(A,X) & = & g_k(A,X) \end{array} \qquad (1)$$

where

$X = \{x_1,x_2...x_n\}$ are known variables

$A = \{a_1,a_2...a_m\}$ are unknown variables

From this equation set, a tautology for the system is given as follows:

---

*This work was supported by the U. S. Army Research Office.

$$
\begin{aligned}
f_1 g_1 \vee \bar{f}_1 \bar{g}_1 &= \psi_1(A,X) \equiv 1 \\
&\vdots \\
f_k g_k \vee \bar{f}_k \bar{g}_k &= \psi_k(A,X) \equiv 1 \\
\psi_1 \cdot \psi_2 \cdot \quad \ldots \psi_k &= \phi\ (A,X) \equiv 1
\end{aligned}
\tag{2}
$$

The desired solution is

$$
\begin{aligned}
A_1 &= h_1(X) \\
&\vdots \\
A_m &= h_m(X) \text{ such that } \phi\ (A,X) \equiv 1
\end{aligned}
\tag{3}
$$

In order to show the solution some notational preliminaries must be defined.

The variable vector A may be described in terms of its atoms as

$$A = A^{\cdot}\ E$$

where

$$A^{\cdot} = \{A^0, A^1, \ldots\ A^{2^m-1}\} \tag{4}$$

and

$$E = \begin{bmatrix} 0 & 0 & \ldots. & & 0 \\ 0 & 0 & \ldots. & 0 & 1 \\ 0 & 0 & \ldots. & 1 & 0 \\ & & \vdots & & \\ 1 & 1 & \ldots. & 1 & 1 \end{bmatrix} \quad 2^m \text{x} m \tag{5}$$

The atoms of A and X of equation (3) are related by a one-zero transformation matrix

$$A^{\cdot} = X^{\cdot} H \tag{6}$$

hence A is related to X by

$$A = X^{\cdot} HE \tag{7}$$

The equations of (1) written in disjunctive form are

$$\sum_{\alpha} (A,X)^{\alpha}\ f_i(\alpha) = \sum_{\beta} (A,X)^{\beta}\ g_i(\beta) \tag{8}$$

which may be given in a shorthand form as

$$(A,X)^{\alpha}\ f_{i\alpha} = (A,X)^{\beta}\ g_{i\beta} \tag{9}$$

The tautology of equation (2) is then

$$\phi(A,X) = A^{\alpha}\ X^{\beta}\ \phi_{\alpha\beta} \tag{10}$$

The desired solution (equation (3)) is

$$A^{\alpha} = X^{\gamma}\ H^{\alpha}_{\gamma}\ . \tag{11}$$

Substituting eq. (11) into eq. (10) and noting that atoms are orthogonal, i.e.,

$$X^{\gamma} X^{\beta} = \left\{ \begin{matrix} X^{\beta} \text{ if } \beta = \gamma \\ 0 \text{ otherwise} \end{matrix} \right\}$$

yields

$$\phi(A,X) = X^{\beta}\ \phi_{\alpha\beta} H^{\alpha}_{\beta} \equiv 1 \tag{12}$$

The problem then is to find an H such that the tautology holds.

A short example will be given to demonstrate the approach. Suppose the tautology of an equation set is given by

$$(x_1\bar{x}_2 + \bar{x}_1 x_2)\, a_1 a_2 + (x_1 x_2 + \bar{x}_1)\bar{a}_1\bar{a}_2 \equiv 1 \qquad (13)$$

In canonical form this is

$$(x^2 + x^1)\, a^3 + (x^3 + x^1 + x^0)\, a^0 \equiv 1$$

and $\phi_{\alpha\beta}$ may be shown as

$$\phi_{\alpha\beta} \Rightarrow \begin{array}{c|cccc} \beta \diagdown \alpha & 0 & 1 & 2 & 3 \\ \hline 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 \\ 3 & 1 & 0 & 0 & 0 \end{array}$$

For a solution to exist there must be at least one entry per row (which is true here). In addition, the solution

$$A^{\cdot} = X^{\cdot}H$$

must have one and only one atom of A associated with an atom of X. Hence, there are two allowable H matrices for this example,

$$H_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and } H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

which give solutions for A of

$$\begin{array}{cccc} A & = X^{\cdot} & H & E \\ [a_1, a_2] & = [x^0, x^1, x^2, x^3] & \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \end{array}$$

$$a_1 = a_2 = x_1\,\bar{x}_2$$

and also

$$a_1 = a_2 = x_1\bar{x}_2 + \bar{x}_1 x_2$$

The task at hand now is to formulate specific design problems in a way in which the desired information can be derived from a solution.

3. APPLICATION OF EQUATIONS

The first application to be considered is that of decomposition. This problem was treated in part by Ledley [5]. The second application considered is the formulation of the cascade [6] design for a specific function. This is covered in detail with an example. Finally, the constrained input problem [2] is formulated using the approach of the previous section.

A pictorial view of decomposition is given in Figure 1. The equation set required is

$$z = f(X) = g(A)$$

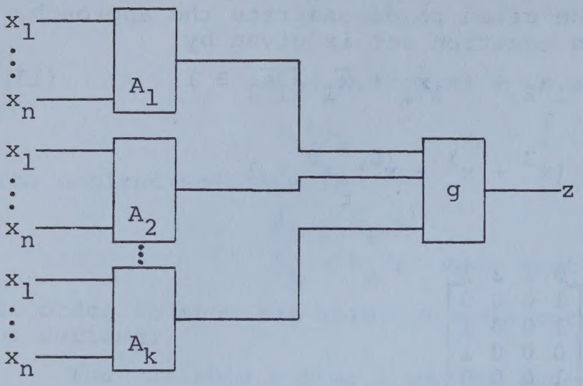


Figure 1. Network Decomposition

f and g are defined by the specific problem and the desired solution is given by

$$A = X^{\cdot}\ H\ E.$$

The solution will include all possible decompositions of k or fewer variables. The equation set may be expanded for multiple levels of decomposition, e.g., constrained input multi-level NAND synthesis. Each additional level requires a defining equation and further constrains the allowable solutions.

The decomposition problem deals with a constrained topology and constrained connectives. The array of Figure 2 may be considered to be a fixed topology, however, the connective of any cell can be variable (either AND, OR, or EOR). The design for this problem amounts to finding the required connectives, if such a set exists. The u variables ($u_{11}\ u_{12}\ u_{21}\ u_{22}$) are used to select the desired cell function. The equation set is

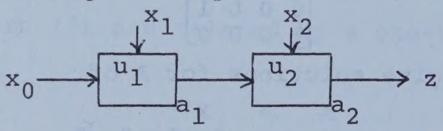


Figure 2. Cascade

$$z = f(x) = u_2^0(x_2 a_1) + u_2^1(x_2+a_1) + u_2^2\ (x_2 \oplus a_1)$$

$$a_1 = u_1^0\ (x_0 x_1) + u_1^1\ (x_0+x_1) + u_1^2\ (x_0 \oplus x_1)$$

The x variables are known and u, A variables unknown. The desired solution is

$$u = X^{\cdot}\ H\ E$$

where u must be a constant (not a function of x). An example will now be given for a specific problem, i.e., with f(x) defined.

Example:
with $z = f(x) = x_0 x_1 \bar{x}_2 + \bar{x}_0 x_2 + \bar{x}_1 x_2$
and the other equations as before, the resulting ϕ matrix is given in Figure 3. The columns (Fig. 3) marked with d are don't cares for this problem. First of all no solution exists if there is a row without at least one-one. Here the only solution which gives u as a constant is

$$u_{11}\ u_{12}\ u_{21}\ u_{22} = 0\ 0\ 1\ 0$$

which means that an AND is required in cell one and EOR in cell two.

$$
\phi = \begin{array}{rr*{32}{c}}
 & & \multicolumn{10}{l}{a_1u_{11}u_{12}u_{21}u_{22}} \\
 & & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 & 30 & 31 \\
x_0x_1x_2 & 0 & 1 & 1 & 1 & & 1 & 1 & 1 & & 1 & 1 & 1 \\
 & 1 & & 1 & 1 & & & 1 & 1 & & & 1 & 1 \\
 & 2 & 1 & 1 & 1 & & & & & & & & & & & & & & & & & & 1 & & & & 1 \\
 & 3 & & 1 & 1 & & & & & & & & & & & & & & & & & & 1 & 1 & & & 1 & 1 \\
 & 4 & 1 & 1 & 1 & & & & & & & & & & & & & & & & & & 1 & & & & 1 \\
 & 5 & & 1 & 1 & & & & & & & & & & & & & & & & & & 1 & 1 & & & 1 & 1 \\
 & 6 & & & & & & & & & & & & & & & & & & 1 & 1 & & & 1 & 1 \\
 & 7 & & & & & & & & & 1 & & & & & & & & & & 1 & & & & 1 \\
 & & & & & d & & & & d & & & & d & d & d & d & d & & & & d & & & & d & & & & d & d & d & d & d
\end{array}
$$

Figure 3. ø Matrix for Cascade

The final application to be covered here is that of designing a network which has constrained inputs. Such a network is shown in Figure 4.

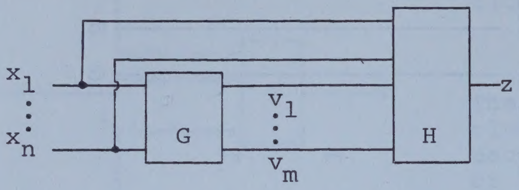


Figure 4. Constrained Input Problem

Here the desired z(x) is defined and may be realized with x variables as well as the constrained variables V(x) which are also available and defined. The defining equations are

$$z = f(x) \rightarrow \psi_1 \ (z,x)$$

$$v_i = g_i(x) \rightarrow \psi_i \ (x)$$

The final tautology $\phi(z,x)$ is the union of the ψ's rather than intersection since the V equations are optional. Here X, V are treated as known variables and z as the unknown. The desired solution is

$$z = (X,V) \cdot HE$$

The valid solutions will include z(X), z(X,V) and z(V) if any exist and it is not necessary to treat these as separate cases.

4. SUMMARY AND EXTENSIONS

The solution method presented provides a method for handling a very general and flexible set of Boolean equations. It is, however, not always obvious just how to formulate a given problem in a way which will provide meaningful solutions.

One of the major difficulties encountered in problems which have been solved is that a great many possible solutions are generated. Recognizing the 'good' solutions is sometimes difficult. It would be desirable to be able to write additional constraint equations which limit the solution space.

This approach is difficult to apply using paper and pencil and it is hoped that continued experience with computer aided solutions will give the insights desired.

REFERENCES

1. Brown, F. M., "Reduced solutions of Boolean equations", Trans. IEEE, Vol. C-19, pp. 976-981, Oct. 1970.
2. ———————, "The constrained-input problem", Proc. 1973 Southeastern Conf., 1973.
3. Gault, J. W., "On solving Boolean equations in n unknowns", Proc. Mexico Int'l. Conf., pp. 408-412, 1971.

4. Hammer, P. L., and S. Rudeanu, Boolean Methods in Operations Research, New York, Springer-Verlag, 1968.

5. Ledley, R. S., Digital Computer and Control Engineering, McGraw-Hill, 1960.

6. Maitra, K. K., "A map approach to the solution of a class of Boolean functional equations", Trans. AIEE, Vol. 81, Part I, pp. 34-36, March 1962.

A COST ANALYSIS OF HAZARD-FREE NETWORKS SYNTHESIZED FROM MONOTONE JUMP FUNCTIONS

E. G. DUCASSE
Department of Computer and
Information Science
Brooklyn College, CUNY
Brooklyn, New York 11210

G. A. METZE
Department of Electrical Engineering
and Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

ABSTRACT

Recently several ways have been found to implement Boolean functions in forms free of static and dynamic hazards by using functions defined on Post algebras [4,11]. In the cost analysis part of this work, however, the Post functions employed were multiple-threshold devices, which are not easily implemented physically. The functions treated in this paper, on the other hand, are either monotone increasing or monotone decreasing and make only a single change in signal value. They can therefore be implemented by single-threshold devices. This paper shows how hazard-free realizations of Boolean operators can be constructed from these single-threshold functions. A number of these implementations are considered for each Boolean operator, and a cost analysis is done for these networks to determine a least expensive hazard-free form for each function.

I. INTRODUCTION

This paper presents a method of solving a common problem in binary switching theory, the static and dynamic hazard-free realization of functions defined on Boolean algebras, by employing functions defined on structures called Post lattices, a generalization of Boolean algebras [6,9]. The Post functions used in these realizations must, however, be free from operation hazards [11], a type of temporary false output unique to multivalued devices. The main emphasis of the paper is on obtaining not only a hazard-free implementation of a given Boolean operator using Post functions, but on a least expensive such realization. First an outline will be given of a particular way in which a Boolean operator can be expressed in hazard-free form by means of Post functions. From this single expression, a number of other hazard-free forms will be derived. Finally, a cost comparison will be made to determine the least expensive such realization.

We begin by outlining a way in which a Boolean operator can be expressed in a hazard-free form in terms of Post functions. Let $f(x_1,\ldots,x_k) \in F[B(2^n) : x_1,\ldots,x_k]$, the class of all k-variable Boolean operators defined on the Boolean algebra of 2^n elements. The first step in obtaining a static and dynamic hazard-free realization of $f(x_1,\ldots,x_k)$ is simply to write it as a composition of one-variable functions defined on the same lattice $B(2^n)$. A good way to do this is to use the canonical sum-of-products form. The one-variable Boolean operators $f(x_i) \in F[B(2^n) : x_i]$ are then mapped by an isomorphism onto $G_B[P(2n):x_i]$, the class of those elements in the Post lattice of one-variable operators defined on the 2n-chain P(2n) which have complements. The lattice $G[P(2n):x_i]$ is isomorphic to $[P(2n)]^{2n}$, the direct product of 2n copies of P(2n). This means that each function in $G[P(2n):x_i]$ can be represented as a 2n-tuple, each of whose components ranges between 0 and 2n-1.

Next a certain relation is defined in $G[P(2n):x_i]$: two elements f and g of this algebra are defined to be equivalent, denoted $f \sim g$, if they have zeros for precisely the same components in their 2n-tuple representations. This can easily be shown to be an equivalence relation, and it can be demonstrated that each equivalence class determined by this relation contains a unique element of $G_B[P(2n):x_i]$. The important feature of the way this relation partitions $G_B[P(2n):x_i]$ is that each equivalence class also contains a unique function, all of whose values are either 0 or 1. These functions can then be used in place of the members of $G_B[P(2n):x_i]$ in the representation of $f(x_1,\ldots,x_k)$. The significance of this replacement is that each of the operators assuming only the values 0 and 1 can be implemented by a circuit free of static and dynamic hazards [11]. Hence, $f(x_1,\ldots,x_k)$ also has such an implementation.

Thus every Boolean function defined on a finite Boolean algebra can be realized by an asynchronous network which is free of static and dynamic hazards. Employing the method outlined above, one obtains a unique representation of a given operator. In the next section it will be shown how additional representations can be obtained. We will then investigate how all the Post functions employed in these realizations can be constructed using monotone functions which have only a single change in signal value. Finally, the costs of these various hazard-free implementations of a Boolean operator will be compared.

II. FUNDAMENTAL CONCEPTS

Since the concept of Post algebra may not be familiar, we begin by stating for reference its definition.

DEFINITION 1. Let n be an integer satisfying $n \geq 2$. A Post algebra P is a distributive lattice with zero element θ and unit element U for which the following conditions hold:

AXIOM 1. There exist n elements $e_o,\dots,e_{n-1}$ in P satisfying

(1a) $\theta = e_o \leq e_1 \leq \cdots \leq e_{n-1} = U$;

(1b) if $x \in P$ and $x\cdot e_1 = \theta$, then $x = \theta$;

(1c) if $x \in P$ and $x + e_{i-1} = e_i$ for some i, then $x = e_i$.

AXIOM 2. For each element $x \in P$, there exist n elements $C_o(x),\dots,$ $C_{n-1}(x)$ in P satisfying

(2a) $C_i(x)\cdot C_j(x) = \theta$ if $i \neq j$;

(2b) $\sum_{i=0}^{n-1} C_i(x) = U$.

AXIOM 3. For each $x \in P$, $x = \sum_{i=0}^{n-1}(e_i \cdot C_i(x))$.

It is clear from this definition that Post algebra is an extension of Boolean algebra: the elements e_i are generalizations of the constants θ and U, and the functions C_i generalize the identity and complement functions on a Boolean algebra. Every Boolean algebra B is a Post algebra with $n = 2$, where $e_o = \theta$, $e_1 = U$, $C_o(x) = \overline{x}$ and $C_1(x) = x$ for all $x \in B$. Note that not every element of a Post algebra P need have a complement. Those elements of P which are complemented form a Boolean algebra, called the underlying Boolean algebra of P and denoted by P_B. It is known [9] that finite Post algebras are precisely the direct products $[P(n)]^m$ of m copies of the chain of n nodes, where $m \geq 1$ and $n \geq 2$ are integers. Elements of such algebras can thus be denoted by n-tuples, each of whose components lies between 0 and n-1. This representation theory will be very important in the sequel.

A Post function is defined on a Post algebra in much the same way that a Boolean function is defined on a Boolean algebra. If $x_1,\dots,x_k$ are variables whose common domain is a Post algebra P, then a function $g(x_1,\dots,x_k)$ built up from these variables and from elements of P by a finite number of applications of the operations +, ·, and $C_o,\dots,C_{n-1}$, is called a Post function of the variables $x_1,\dots,x_k$. The Post functions of interest in this paper are those which have implementations which produce no temporary false signal values. To introduce this concept, consider a k-tuple $v = v_1\dots v_k$, each of whose components v_j satisfies $0 \leq v_j \leq n-1$. Such vectors can be considered to be the k-tuple arguments of operators $g(x_1,\dots,x_k)$ defined on the Post chains of $n \geq 2$ nodes; that is, of members of the Post algebra $G[P(n):x_1,\dots,x_k]$. Two k-tuples $u = u_1\dots u_k$ and $v = v_1\dots v_k$ satisfying $0 \leq u_j,v_j \leq n-1$ for all $j = 1,\dots,k$ are said to be adjacent if $|u-v| = \sum_{j=1}^{k}|u_j-v_j| \leq 1$. Employing this concept, the notion of operation hazard can be defined for Post functions.

DEFINITION 2. If u and v are adjacent argument vectors of the Post function defined on P(n) and $|g(u)-g(v)| \geq 2$, then $g(x_1,\ldots,x_k)$ is said to contain an operation hazard. If $|g(u)-g(v)| \leq 1$ for all adjacent vectors u and v, then g is said to be operation-hazard-free, denoted O-H-F.

In an asynchronous circuit implementing a function $g(x_1,\ldots,x_k)$, the output changes from g(u) to g(v) when the input vector changes from u to v. If $|g(u)-g(v)| \geq 2$, signal values other than g(u) and g(v) may appear temporarily at the output of the network during the transition from g(u) to g(v). Therefore any asynchronous circuit realizing the function g will produce an operation hazard. This, together with the representation theory for finite Post lattices, explains why the elements of $G_B[P(2n):x_i]$ cannot be used to obtain a hazard-free realization of a Boolean operator. Since $G[P(2n):x_i] \cong [P(2n)]^{2n}$, the elements of $G_B[P(2n):x_i]$ can be regarded as 2n-tuples, each of whose components is either 0 or 2n-1 [2]. In fact, if 0,1,...,2n-1 are the elements of P(2n), then any $g = g_o \cdots g_{2n-1}$ in $G[P(2n):x_1,\ldots,x_k]$ can be defined by $g_j = g(j)$ for all $j = 0,\ldots,2n-1$. Now Definition 2, restricted to one-variable functions g, states that g has an operation hazard if there exist $u,v \in P(2n)$ satisfying $|u-v| \leq 1$ and $|g(u)-g(v)| \geq 2$. Thus all elements of $G_B[P(2n):x_1,\ldots,x_k]$ except the constant functions g(x) = 0 and h(x) = 2n-1 have operation hazards, because the assumed linear ordering of the signal values 0,...,2n-1 implies that all intermediate values will be traversed in a transition between 0 and 2n-1. The functions, all of whose values are 0 or 1, on the other hand, are O-H-F, and thus have implementations free from static and dynamic hazards [11]. Also, combining them by means of + and · gates yields no operation hazards, demonstrating that a Boolean operator expressed as a composition of these functions can be implemented in a hazard-free form.

The functions defined on P(n) which assume only the values 0 and 1 comprise the interval $[e_o,e_1]$ in $G[P(n):x]$ [2]. It follows from the n-tuple representation that the set of all one-variable functions defined on P(n), all of whose values are i or i+1, $0 \leq i \leq n-2$, is precisely the interval $[e_i,e_{i+1}]$. Since it is known [4] that the function φ defined for $g \in [e_o,e_1] \subseteq G[P(n):x]$ by $\varphi(g) = e_i + e_{i+1} \cdot C_o^2(g)$ is an isomorphism of $[e_o,e_1]$ onto $[e_i,e_{i+1}]$ for all $i = 0,\ldots,n-2$, it follows that any interval of functions $[e_i,e_{i+1}]$, $0 \leq i \leq 2n-2$, in $G[P(2n):x_i]$ may be used as an alternative to $[e_o,e_1]$ in the hazard-free realization of Boolean operators on $B(2^n)$.

The sets of mappings $[e_i, e_{i+1}]$ in $G[P(2n):x_i]$ are thus seen to be distinguished classes of this lattice. Indeed, they possess additional desirable properties. If, in the hazard-free implementation of Boolean functions, we consider only those sets of Post operators which already contain all the elements necessary to realize any Boolean function on $B(2^n)$ so that no additional transformations need be synthesized, then the intervals $[e_i, e_{i+1}]$ are the minimal such classes. In fact, among such classes, these intervals are the only collections which should be employed in the implementation of Boolean operators; for any other class, even though it may consist solely of O-H-F elements, will always contain at least two functions which when gated by $+$ or $\cdot$, give rise to operation hazards. Now that we have seen how any of the 2n-1 intervals of O-H-F functions $[e_i, e_{i+1}]$ in $G[P(2n):x_i]$ can be used in the hazard-free realization of Boolean operators, a cost analysis will be done to determine the most desirable interval.

III. THE COST ANALYSIS

An earlier cost analysis for hazard-free realizations of Boolean operators [4] considered realizations in terms of the functions C_i. However, these functions are difficult to implement physically. The operators we introduce now, on the other hand, present no such problem since they can be implemented by single-threshold devices. For $i = 1,\ldots, 2n-2$ we define functions $g_i(x)$ on $P(n)$ by $g_i(x) = \underbrace{(n-1)\ldots(n-1)}_{i}\ \underbrace{0\ldots0}_{n-1}$ if $1 \leq i \leq n-1$, and $g_i(x) = \underbrace{0\ldots0}_{i-n+1}\ \underbrace{(n-1)\ldots(n-1)}_{2n-i-1}$ if $n \leq i \leq 2n-2$. Since all of these 2n-2 functions are either monotone increasing or monotone decreasing and each has a single change in signal value, we call them monotone jump functions. It is known [11] that this set of operators generates every Post function on $P(n)$.

If we wish to be able to realize any k-variable Boolean operator on $B(2^n)$ in a form free of static and dynamic hazards, it will be necessary to have available all the members of some interval $[e_i, e_{i+1}]$ of the Post algebra $G[P(2n):x_i]$. This implies that a cost comparison should be carried out for entire intervals of O-H-F Post functions. Fortunately, all these operators can be put into an easily recognizable standard form from which cost comparisons can be made. Our cost analysis will be carried out in an arbitrary lattice $G[P(n):x] \cong [P(n)]^n$. Assume $g \in [e_i, e_{i+1}] \subseteq G[P(n):x]$. The cost criterion we adopt for these Post functions is a straightforward generalization of the standard criterion for Boolean operators. It is the total number of input leads to all logic elements in a network. In this criterion it will be assumed that the C_i operators and constant functions are not available as inputs, but rather will cost one lead each.

Each function $g \in [e_i, e_{i+1}]$ can be regarded as being a finite sequence of blocks of i's and (i+1)'s. Assume the number of blocks containing the initial value in the n-tuple representation $g = g_o \cdots g_{n-1}$ is ℓ. We will derive a standard form for g depending solely on its block structure. There are four distinct cases: (1) $B_{i_1} B_{(i+1)_1} B_{i_2} B_{(i+1)_2} \cdots B_{i_{(\ell-1)}} B_{(i+1)_{(\ell-1)}} B_{i_\ell}$ - g begins and ends with i; (2) $B_{i_1} B_{(i+1)_1} \cdots B_{i_\ell} B_{(i+1)_\ell}$ - g begins with i and ends with i+1; (3) $B_{(i+1)_1} B_{i_1} \cdots B_{(i+1)_\ell} B_{i_\ell}$ - g begins with i+1 and ends with i; and (4) $B_{(i+1)_1} B_{i_1} \cdots B_{(i+1)_{(\ell-1)}} B_{i_{(\ell-1)}} B_{(i+1)_\ell}$ - g begins and ends with i+1. To find a standard form for the functions of types (1) and (2), let $j_m = \sum_{p=1}^{m} |B_{i_p}| + \sum_{p=1}^{m-1} |B_{(i+1)_p}| + 1$ and $k_m = \sum_{p=1}^{m} |B_{i_p}| + \sum_{p=1}^{m} |B_{(i+1)_p}|$. The standard form we adopt for case (1) is $i \dot{+} (i+1) \cdot [\sum_{m=1}^{\ell-1} g_{k_m} \cdot g_{n+j_m-2}]$, where $\dot{+}$ denotes addition for real numbers. For case (2), the standard form is $i \dot{+} (i+1) \cdot [\sum_{m=1}^{\ell-1} g_{k_m} \cdot g_{n+j_m-2} + g_{n+j_\ell-2}]$. For cases (3) and (4), let $r_m = \sum_{p=1}^{m-1} |B_{i_p}| + \sum_{p=1}^{m-1} |B_{(i+1)_p}| + 1$ and $s_m = \sum_{p=1}^{m-1} |B_{i_p}| + \sum_{p=1}^{m} |B_{(i+1)_p}|$. The standard form for case (3) is $i \dot{+} (i+1) \cdot [g_{s_1} + \sum_{m=2}^{\ell} g_{s_m} \cdot g_{n+r_m-2}]$, while for case (4) it is $i \dot{+} (i+1) \cdot [g_{s_1} + g_{n+r_\ell-2} + \sum_{m=2}^{\ell-1} g_{s_m} \cdot g_{n+r_m-2}]$, except when g is the constant function, all of whose values are i+1. In this case, $g_{s_1} + g_{n+r_\ell-2} = U$, so the standard form is merely i+1.

Our next goal is to determine the cost of a function $g \in [e_i, e_{i+1}]$ in its standard form. In case (1), if $\ell = 1$, then we have the constant function $g \equiv i$, whose cost is 1 for all i = 0,...,n-2. If $\ell = 2$ and $0 < i < n-2$, the cost of g is 9, while for $\ell \geq 3$ and $0 < i < n-2$ it is $5\ell + 1$. If $i = 0$ and $\ell = 2$, the cost of g is 6, while for $\ell \geq 3$ it is $5\ell - 2$. Finally, if i = n-2 the cost is 7 regardless of the value of $\ell \geq 2$. Next consider case (2). If $0 < i < n-2$ and $\ell = 1$, the cost of g is 7, while if $\ell \geq 2$ it is $5\ell + 3$. If $i = 0$ and $\ell = 1$, the cost is 4, whereas for $\ell \geq 2$ it is 5ℓ. Finally, for i = n-2 the cost is $5\ell - 1$ independent of the value of $\ell \geq 1$. Turning to case (3), if $0 < i < n-2$ and $\ell = 1$, the cost of g is 7, while if $\ell \geq 2$ it is $5\ell + 3$. If $i = 0$ and $\ell = 1$, the cost is 4, whereas for $\ell \geq 2$ it is 5ℓ. Finally, for i = n-2 the cost is $5\ell - 1$ for all $\ell > 1$. Note that the costs in case (3) are the same as the corresponding costs in case (2).

Now considering case (4), if $\ell = 1$ we have the constant function $g \equiv i+1$, whose cost is 1 for all $i = 0,\dots,n-2$. If $0 < i < n-2$ and $\ell = 2$, then the cost of g is 10, while if $\ell \geq 3$ it is $4\ell+3$. If $i = 0$ and $\ell = 2$, the cost is 7, whereas for $\ell \geq 3$ it is 4ℓ. Finally, for $i = n-2$ and $\ell = 2$, the cost is 6, while for $\ell \geq 3$ it is $4\ell-1$. All these results are summarized in Table 1. It can be seen from this table that every function has a cheapest realization in $[e_{n-2}, U]$, except that those of the form $B_{0_1} B_{1_1} B_{0_2}$ have one less input lead than those of the form $B_{(n-2)_1} B_{(n-1)_1} B_{(n-2)_2}$. We therefore conclude that the interval $[e_i, e_{i+1}]$ of Post functions from $G[P(2n):x]$ which should be chosen for the hazard-free implementation of Boolean operators from $F[B(2^n):x_1,\dots,x_k]$ is $[e_{n-2}, U]$.

IV. CONCLUSION

It has been shown that there is always more than one way of expressing a Boolean operator in a hazard-free form in terms of monotone jump functions. The costs of these various implementations have been evaluated and compared, and one particular set of functions has been determined to be the least expensive in every case but one. This set then is recommended for use in the hazard-free implementation of Boolean operators.

ℓ	Case	$i = 0$	$0 < i < n-2$	$i = n-2$
$\ell = 1$	(1)	1	1	1
	(2)	4	7	4
	(3)	4	7	4
	(4)	1	1	1
$\ell = 2$	(1)	6	9	7
	(2)	10	13	9
	(3)	10	13	9
	(4)	7	10	6
$\ell \geq 3$	(1)	$5\ell-2$	$5\ell+1$	$5\ell-3$
	(2)	5ℓ	$5\ell+3$	$5\ell-1$
	(3)	5ℓ	$5\ell+3$	$5\ell-1$
	(4)	4ℓ	$4\ell+3$	$4\ell-1$

TABLE 1

REFERENCES

1. Birkhoff, G., Lattice Theory, American Mathematical Society Colloquium Publication 25, revised edition, New York, 1948.

2. DuCasse, E. and Metze, G., "Some Results on Computing Function Values in Finite Post Algebras," Coordinated Science Laboratory Report R-539, University of Illinois, Urbana, 1971.

3. DuCasse, E. and Metze, G., "The Relation of Equivalence for Post Algebras," Proceedings of the Tenth Annual Allerton Conference on Circuit and System Theory, Urbana, Illinois, pp. 625-634, October, 1972.

4. DuCasse, E. and Metze, G., "Hazard-Free Realizations of Boolean Functions Using Post Functions," Conference Record of the 1973 Symposium on the Theory and Applications of Multiple-Valued Logic Design, Toronto, Canada, pp. 59-67, May, 1973.

5. DuCasse, E. and Metze, G., "Comments on 'The Relationship Between Boolean Algebra and Multivalued Switching Algebra Under Different Definitions of Complement'," to be published in the IEEE Transactions on Computers.

6. Epstein, G., "The Lattice Theory of Post Algebras," Transactions of the American Mathematical Society, Vol. 95, pp. 300-317, 1960.

7. Metze, G., "An Application of Multivalued Logic Systems to Circuits," Proceedings of a Symposium on Circuit Analysis, University of Illinois, Urbana, pp. 11-1 to 11-14, 1955.

8. Post, E., "Introduction to a General Theory of Elementary Propositions," American Journal of Mathematics, Vol. 43, pp. 163-185, 1921.

9. Rosenbloom, P. C., "Post Algebras. I. Postulates and General Theory," American Journal of Mathematics, Vol. 64, pp. 167-188, 1942.

10. Trazcyk, T., "Axioms and Some Properties of Post Algebras," Colloquium Mathematicum, Vol. 10, pp. 193-209, 1963.

11. Wojcik, A. S., "Relationships Between Post and Boolean Algebras with Application to Multivalued Switching Theory," Coordinated Science Laboratory Report R-512, University of Illinois, Urbana, 1971.

12. Wojcik, A. S. and Metze, G., "Some Relationships Between Post and Boolean Algebras," Conference Record of the 1971 Symposium on the Theory and Applications of Multiple-Valued Logic Design, Buffalo, New York, pp. 173-182, May, 1971.

ON THE DIAGNOSIS OF MULTIPLE OUTPUT CIRCUITS

DONALD R. SCHERTZ
Professor of Electrical Engineering
Bradley University, Peoria, Illinois

ABSTRACT

The fault diagnosis of multiple output combinational circuits is considered in terms of a circuit representation based on classes of structurally equivalent faults. The use of a vector notation allows multiple output circuits to be treated in the same manner as single output circuits.

INTRODUCTION

One approach to the diagnosis of digital circuits utilizes some transformed version of the circuit including all possible faults as a basis for the derivation of test sequences. The particular transformation under consideration here is based on classes of structurally equivalent faults. This transformation of the circuit has been presented previously for the case of single output combinational circuits [1]. In this paper, the work is extended to include multiple outputs. A brief review of the transformation procedure as applied to single output circuits is presented as the basis for the extension to multiple output circuits. This is followed by a discussion of the vector Boolean operation to be used and by examples of the application to fault detection and fault location.

TRANSFORMED OUTPUT EXPRESSION

In the case of single output circuits, the circuit under consideration is represented by a single Boolean expression, the transformed output expression $ZT(\tilde{X},\tilde{F})$ involving variables representing each of the primary inputs of the original circuit and each of the classes of structurally equivalent faults. The arguments of the function are the input vector $\tilde{X} = X_1,X_2...X_m$ where the input variables $X_1,X_2...X_m$ represent the primary inputs to the circuit and the fault vector $\tilde{F} = F_1,F_2...F_n$ where the fault variables $F_1,F_2...F_n$ represent the occurrence of faults in the various classes. The fault variable F_i takes on the value 1 if and only if the circuit contains one of the faults in fault class i. $ZT(\tilde{X},\tilde{F})$ is constructed by means of an iterative procedure starting at the fault class closest to the primary output and working back to the primary inputs. This procedure is algorithmic and is easily carried out either by hand computation or by means of a digital computer program. The reader is referred to the previously mentioned reference for details [1].

The extension to multiple output circuits is made by forming a vector transformed output expression as shown in Equation 1.

$$\widetilde{ZT}(\tilde{X},\tilde{F}) = ZT_1(\tilde{X},\tilde{F}),\ ZT_2(\tilde{X},\tilde{F}),\cdots ZT_K(\tilde{X},\tilde{F}) \qquad (1)$$

Each component of this vector expression is the transformed output expression for one of the output lines of the circuit under consideration. Note that each component of $ZT(\tilde{X},\tilde{F})$ is considered to be a function of all m input lines and all n fault classes, even though a given component might well be vacuous in one or more of the input or fault variables.

The vector transformed output expression $\widetilde{ZT}(\tilde{X},\tilde{F})$ can be used to describe circuit behavior under specified fault or input conditions. For example, the behavior of the circuit in the presence of a fault described by fault vector $\tilde{F}_j$ is found by evaluating $\widetilde{ZT}(\tilde{X},\tilde{F}_j)$. Similarly, the behavior of the fault free circuit is obtained by substituting the all 0 fault vector to evaluate $\widetilde{ZT}(\tilde{X},\tilde{0})$. In each case, the resulting vector expression has a component for each output line. Each component consists of a Boolean expression which gives the function developed on that output line in terms of the primary input variables.

The behavior of the circuit with a specific input is obtained by evaluating $\widetilde{ZT}(\tilde{X}_i,\tilde{F})$ where $\tilde{X}_i$ specifies the applied input vector. Again the resulting expression contains one component for each output line, but in this case the components are expressed in terms of the fault variables rather than the primary inputs. Finally, the behavior of the circuit under input $\tilde{X}_i$ and fault $\tilde{F}_j$ is described by the expression $\widetilde{ZT}(\tilde{X}_i,\tilde{F}_j)$. Each component of the vector in this case is simply 0 or 1 giving the actual value on the corresponding output line.

VECTOR EXCLUSIVE OR

The EXCLUSIVE OR operation is often used in diagnostic procedures to check whether the output of a faulty circuit is the same as the output of the corresponding fault free circuit. For example, a given test detects a given fault if and only if the output of the faulty circuit with the given test as input is different from the output of the fault free circuit with the same input. In the case of multiple output circuits, it is sufficient for any one output to differ for faulty and fault free behavior. An operation which produces 0 when applied to two identical vectors and 1 when applied to two vectors which differ in at least one place is thus needed. The VECTOR EXCLUSIVE OR defined below is such an operation.

The VECTOR EXCLUSIVE OR $\tilde{\oplus}$ is a binary vector operation which operates on two vectors of the same length and produces a single output of 0 if the vectors are identical and which produces a single output of 1 if the vectors differ in at least one place. For any two K - element vectors $\widetilde{VA}$ and $\widetilde{VB}$, this may be expressed symbolically as shown in Equation 2.

$$\widetilde{VA} \,\tilde{\oplus}\, \widetilde{VB} = (VA_1 \oplus VB_1) + (VA_2 \oplus VB_2) + \cdots (VA_K \oplus VB_K) \qquad (2)$$

The application of the VECTOR EXCLUSIVE OR to the vector transformed output expression for a circuit allows previously derived diagnostic procedures for single output circuits to be extended to include multiple output circuits.

APPLICATIONS TO DIAGNOSIS

The transformed output expression has been used previously to solve a number of problems relative to the diagnosis of single output circuits [2]. In this section, we present the extension of these results to multiple output circuits. In each case, the vector transformed output expression is used with the VECTOR EXCLUSIVE OR to formulate the desired solution.

A test will detect a given fault if at least one output of the faulty circuit differs from the corresponding output of the fault free circuit. Thus fault $\tilde{F}_j$ is detected by test $\tilde{X}_i$ if Equation 3 is satisfied.

$$\widetilde{ZT}(\tilde{X}_i,\tilde{0}) \,\tilde{\oplus}\, \widetilde{ZT}(\tilde{X}_i,\tilde{F}_j) = 1 \qquad (3)$$

The determination of a suitable test to detect a given fault is readily made by considering Equation 4. Since $\tilde{0}$ and $\tilde{F}_j$ are completely specified binary vectors, the resulting equation involves only the primary inputs $X_1 \cdots X_m$.

$$\widetilde{ZT}(\tilde{X},\tilde{0}) \,\tilde{\oplus}\, \widetilde{ZT}(\tilde{X},\tilde{F}_j) = 1 \qquad (4)$$

The solution of Equation 4 results in the set of all tests which detect the given fault. If no solution to Equation 4 exists, then the fault represented by $\tilde{F}_j$ cannot be detected. The presence of an undetectable fault implies that the original circuit was redundant [3]. It should be noted that the procedure used is the same whether the fault considered is a single or a multiple fault.

A related problem in fault detection involves the determination of which faults are detected by a given test. The appropriate form is given in Equation 5 where $\tilde{X}_i$ represents the input assignment corresponding to the given test.

$$\widetilde{ZT}(\tilde{X}_i,\tilde{0}) \,\tilde{\oplus}\, \widetilde{ZT}(\tilde{X}_i,\tilde{F}) = 1 \qquad (5)$$

In this case, the left side of the equation reduces to an expression involving only the fault variables $F_1 \cdots F_n$. This expression specifies all single and multiple faults detected by the given test.

The vector transformed output expression can also be applied to problems of fault location. The solution to Equation 6 gives the set of all tests which can distinquish between fault $\tilde{F}_j$ and fault $\tilde{F}_K$.

$$\widetilde{ZT}(\tilde{X},\tilde{F}_j) \; \tilde{\oplus} \; \widetilde{ZT}(\tilde{X},\tilde{F}_K) = 1 \qquad (6)$$

In a similar manner, the solution to Equation 7 gives the set of all faults which can be distinquished from fault $\tilde{F}_j$ by test $\tilde{X}_i$.

$$\widetilde{ZT}(\tilde{X}_i,\tilde{F}_j) \; \tilde{\oplus} \; \widetilde{ZT}(\tilde{X}_i,\tilde{F}) = 1 \qquad (7)$$

CONCLUSIONS

The vector transformed output expression has been shown to be a convenient basis for the representation of a number of problems relative to the detection and location of faults in multiple output combinational digital circuits. The techniques presented here are intended primarily as tools to be used in developing a theoretical basis for fault diagnosis. It is hoped that such a theoretical basis will eventually lead to improved techniques for the solution of practical problems in fault diagnosis.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under Grant GK27871.

REFERENCES

1. D. R. Schertz and G. Metze, "A New Representation for Faults in Combinational Digital Circuits," IEEE Trans. on Computers, Vol C-21, pp. 856-866, August 1972.
2. D. R. Schertz, "Applications of a Fault Class Representation to Diagnosis," 1973 International Symposium on Fault Tolerant Computing, June, 1973.
3. J. P. Hayes, "A NAND Model for Fault Diagnosis in Combinational Logic Networks," IEEE Trans. on Computers, Vol. C-20, pp. 1496-1506, December, 1971.

ON THE DETECTION OF BRIDGE AND SHORTED DIODE FAULTS*

SUDHAKAR M. REDDY
Dept. of Electrical Engineering, University of Iowa, Iowa City, Iowa 52242

ABSTRACT

A special test circuit is introduced to facilitate detecting bridge and shorted diode faults in gate networks. Certain bridge faults not included in earlier studies are now considered.

I. INTRODUCTION

Several authors have studied the problem of detecting bridge and shorted diode faults[1-4]. In this paper we will derive techniques to detect a larger class of faults than those considered by the earlier workers[1,2]. In Figure 1 we show some of the faults that were not considered in earlier works. For example, bridge faults α, β, γ, and δ. For the sake of simplicity of arguments we assume positive logic (i.e. leading to AND bridging faults[2]). All the results presented are either true directly or with slight modification of the techniques can be made true for OR type bridge faults. Throughout this paper we will use the notation and terminology developed in references 1 and 2, with respect to faults in combinational gate networks (logic networks constructed from AND, OR, NAND, and NOR gates). The two new fault types to be considered are defined below.

Definition 1: Input Cross Bridge Fault (ICBF) is a bridge fault between two network inputs that are not connected to the same gate. For example, the faults γ and δ shown in Figure 1 are ICBF's.

Definition 2: A Cross Bridge Fault (CBF) is a bridge fault between two leads not in the same path of a network and the fault is not an ICBF. For example, the faults α and β shown in Figure 1 are CBF's.

II. GATE NETWORKS

In this section we initially consider gate networks with at most three levels of logic in which the first level gates (the closest to inputs) are NOT gates if the network has three logical levels. Next gate networks that are tree networks are considered.

Definition 3: Three-level combinational gate networks in which the first level gates are NOT gates if the network has three logic levels, are called restricted three level gate networks (RTLGN's).

---

*The research reported was supported by NSF Grant GK-36377.

Examples of three-level networks that are included and those not included in Definition 3 are given in Figure 2.

The following theorem can be proved for RTLGN's. The proof is not given to conserve space.

Theorem 1: If T is a test set, for a single-output RTLGN, to detect single stuck-at-faults then T detects all detectable single faults of the following types:

(a) Stuck-at-faults,

(b) inverting feedback bridge faults[2] (IFBF's), and

(c) Cross Bridge Faults (CBF's).

Notice that Theorem 1 does not cover ICBF's, IBF's and SDF's. To detect these faults we introduce a special network shown in Figure 3 and call it Special Test Circuit (STC). The important thing to note is that the inputs to STC are the same ones that are applied to a RTLGN under test. Specially if $\bar{x}_i$ is generated in the RTLGN then the same $\bar{x}_i$ is applied as the input to STC. This is illustrated in Figure 4.

Theorem 2: If a single-output RTLGN and STC are simultaneoulsy tested and if no bridge faults between the leads of RTLGN and STC (except of course between network inputs $x_i, \ldots, x_n$ and $\bar{x}_1, \ldots, \bar{x}_n$ and also bridge between a lead in RTLGN and an input) occur then the test set T' will detect all detectable single faults, of the type given below, in the RTLGN and STC, where $T' = T \cup T_1$, and T is a single-s-a-fault-detecting test set for the RTLGN and T_1 is the test set given in Figure 5.

(a) Stuck-at-faults,

(b) IFBF's,

(c) CBF's,

(d) SDF's,

(e) IBF's, and

(f) ICBF's.

Proof: Since clearly T_1 detects single s-a-faults in STC the faults of types (a)-(c) are covered by Theorem 1. Now assume that an ICBF fault has occurred. Since only single faults are assumed to be occurring, without loss of generality, we can assume that the fault is in STC. The four types of ICBF's that need to be considered are:

(i) bridge between x_i and x_j, $i \neq j$,

(ii) bridge between $\bar{x}_i$ and x_j, $i \neq j$,

(iii) bridge between $\bar{x}_i$ and x_i, and

(iv) bridge between $\bar{x}_i$ and $\bar{x}_j$, and $i \neq j$, $1 \leq i, j \leq n$.

Let $t_i \in T_1 (t'_i \in T_1)$ be the input vector in which only ith component is equal to 1(0) and all other components are 0(1). Then type (i) ICBF is detected at 0_1 of STC when t_i or t_j is applied as input; type (ii) and type (iii) ICBF's are detected at 0_2 by the test input in T_1 that has all its elements equal to 1, and; type (iv) ICBF is detected at 0_1 by the test t'_i or t'_j in T_1.

Proof for SDF's and IBF's can be obtained by arguments similar to the ones used above.

Q.E.D.

Next we consider tree networks defined below.

Definition 4: A gate network is called a tree network if the output of a gate is connected to the input of only one gate. The network inputs and their complements are allowed to be fanned out.

Theorem 3: If a tree network and STC are simultaneously tested then $T' = T \cup T_1$, where T detects single s-a-faults in the tree network, will detect all detectable single faults of the type given below:

(i) s-a-faults,
(ii) IBF's,
(iii) SDF's,
(iv) IFBF's, and
(v) ICBF's

Proof: Proof omitted for reasons of space.

III. CONCLUSIONS

We have shown that a special test circuit can be introduced to facilitate testing of shorted diode and bridge faults. The fault class considered has been expanded from those considered in references 1 and 2. Notice that we have not considered non-inverting feedback bridging faults (NIFBF's). These can be detected by applying tests that detect single s-a-faults in a particular order, as shown earlier[2]. It may be necessary to apply some of the s-a-fault detecting tests more than once to detect all detectable NIFBF's.

REFERENCES

[1] A. D. Friedman, "Diagnosis of Short Faults in Combinational Circuits," Digest of 1973 Intern. Symp. on Fault-Tolerant Computing, pp. 95-99, June 1973, Palo Alto, California.

[2] K.C.Y. Mei, "Bridging and Stuck-at Faults," Digest of 1973 Intern. Symp. on Fault-Tolerant Computing, pp. 95-99, June 1973, Palo Alto, California.

[3] H. Y. Chang, "A Method for Digitally Simulating Shorting Input Diode Failures," BSTJ, Vol. 48, pp. 1957-1966, 1966.

[4] J. P. Roth, "Diagnosis of Automata Failures: A Calculus and a Method," IBM Journal, July 1966.

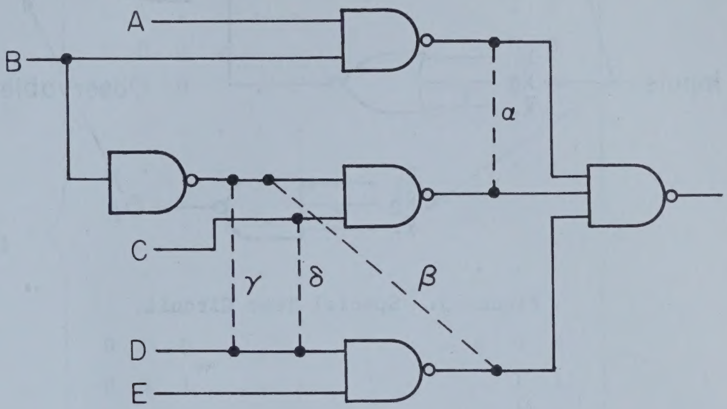
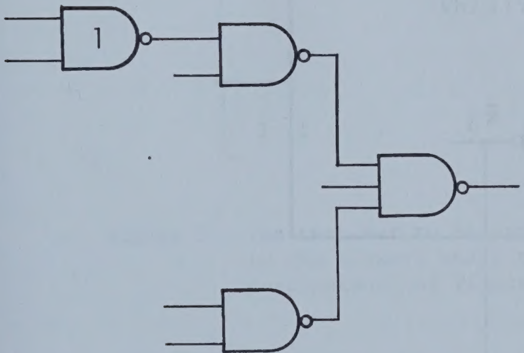
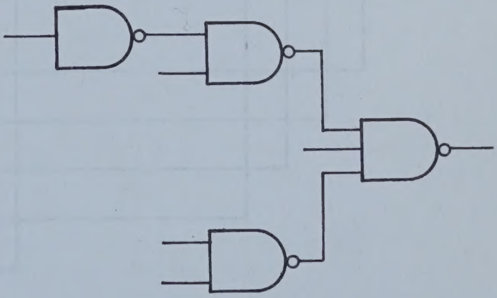


Figure 1. Examples of bridge faults.



(a) Not included since gate 1 is not a NOT gate.



(b) This network is included.

Figure 2. Examples of three level networks included and not included in Definition 3.

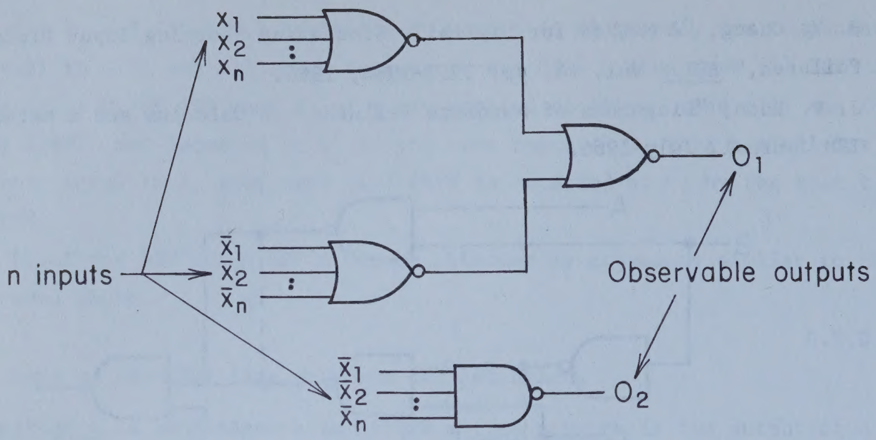

x_1
x_2
x_n
O_1
n inputs
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_n$
Observable outputs
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_n$
O_2

Figure 3. Special Test Circuit.

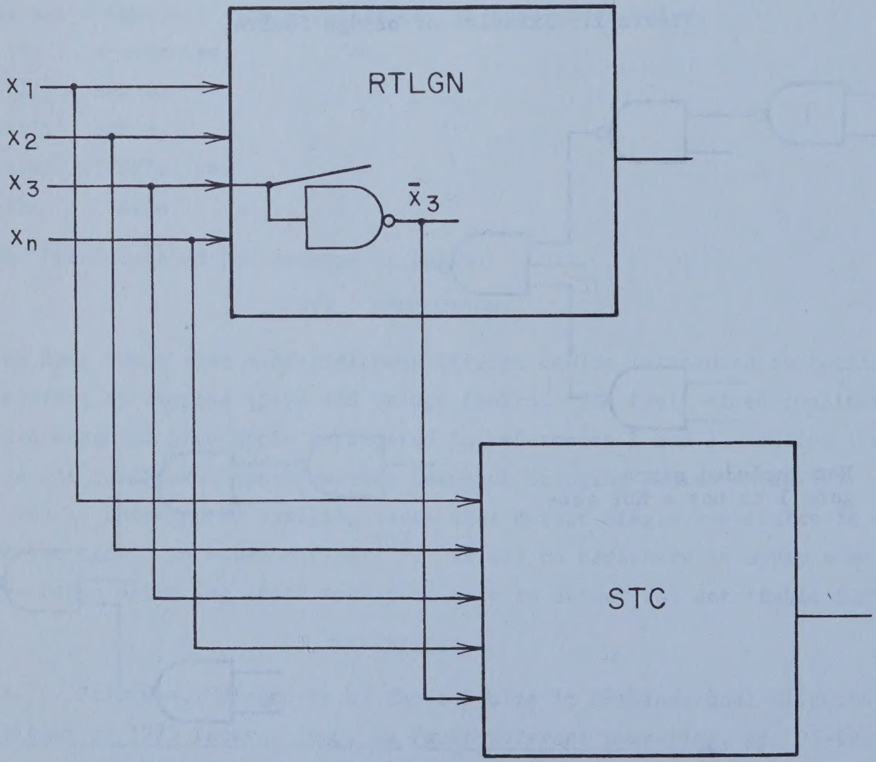

RTLGN
x_1
x_2
x_3
x_n
$\bar{x}_3$
STC

Figure 4. Illustrating inputs to STC.

$$
T_1 = \left\{ \begin{array}{cccccccccc}
x_1 & x_2 & x_3 & x_4 & \cdot & \cdot & \cdot & \cdot & \cdot & x_n \\
\hline
0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 \\
1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 \\
1 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 \\
0 & 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 1 \\
0 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 \\
1 & 0 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & 1 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 0
\end{array} \right\}
$$

Figure 5. The test set to be applied simultaneously to the network under test and the special test network of Figure 3.

AN APPROACH FOR SYNTHESIZING TRANSITION LOGIC CIRCUITS[1]

MISHELL J. STUCKI
Computer Systems Laboratory
Washington University
St. Louis, Missouri

ABSTRACT

A "transition signal" is a change in binary value, regardless of the direction of change. The occurrence of a transition signal is a momentary event similar to the occurrence of a pulse, and these signals can be used in systems design to perform functions traditionally performed by pulses. This paper presents a new approach for synthesizing circuits that accept and generate transition signals.

INTRODUCTION

A "transition signal" is a change in binary level, regardless of the direction of change. If a variable is said to carry transition signals, then a value change from 0 to 1 is a transition signal and so is a value change from 1 to 0. The occurrence of a transition signal is a momentary event similar to the occurrence of a pulse, and these signals can be used in systems design to perform functions traditionally performed by pulses.[2]

Circuits that accept and generate transition signals are called "transition-logic circuits" (TL circuits). The general diagram of such a circuit is shown in Figure 1. Each input and output is a binary variable. Those shown with arrow heads carry transition signals and those shown with diamond heads do not.

A circuit whose behavior is described in terms of transition signals can also be described in terms of levels alone. The level-logic description will usually be sequential because internal states are usually needed to detect and generate transition signals. One method of synthesizing a TL circuit is thus to view it as a level sequential circuit and synthesize it using the Huffman method [4]. There is a problem with this approach, though. The primary flow table that must be constructed is often very large and its construction can be a time consuming and laborious process. It can even be an unmanageable one.

A synthesis technique that does not require a level-logic flow table is thus preferable to the Huffman approach. Chuang [5] has shown that TL circuits can be designed directly from a TL description if one assumes the existance of certain special logic elements. (A special logic element is one not commonly found in commercial lines of logic.) Synthesis methods proposed

---

[1] This work has been supported by the Advanced Research Projects Agency of the Department of Defense under Contract SD-302 and by the Division of Research Resources of the National Institutes of Health under Grant RR-00396.

[2] Transition signals are used extensively in the control networks of the macromodular computer systems [1, 2, 3] under investigation at the Computer Systems Laboratory of Washington University.

by Smith and Roth [6] and by Bredeson and Hulina [7] could also be used even though they were not developed specifically for TL circuits. These methods also depend on special logic elements.

Although the use of special logic elements might make the design of TL circuits easier, it is clear that synthesis techniques based on such elements are of no use to the circuit designer if the brand of logic used in his project does not have these elements. Such methods are thus not ideal alternatives to the Huffman approach. In this paper we present a method for synthesizing TL circuits which we believe to be an ideal alternative. This method works directly from a TL circuit description, generates designs expressed in terms of ordinary combinational operators, and generates solutions in a matter of minutes.

THE CIRCUIT DESCRIPTION

The first step in any synthesis procedure is to express the behavior of the circuit in a specific format. In our case, we require that the circuit behavior be represented by a form of Petri Net [8] which we call a C-net (circuit net). A C-net is composed of circles, bars, directed arcs, and labels. Examples of C-net descriptions are shown in Figures 5, 6, and 7.

A circle is used to represent a condition within the circuit. The condition has two states, assertion and non-assertion. When the state of a condition is to be indicated in a net, a marker (dot) is placed in the circle to indicate assertion.

Certain sets of device conditions are stable and certain sets are not. By definition, an unstable set cannot persist and the occurrence of such a set causes a change in the state of the device. This aspect of device behavior is represented by the bars and by the arcs into and out of the bars. Each bar describes the effect of a specific unstable set. The arcs leading to the bar come from the conditions that make up the set, and when all of these conditions are asserted, the bar is said to "fire". This firing has the following effect: if a circle is connected to the bar by one or more arcs (regardless of the direction of arc), the state of the circle is changed once for each arc. Figure 2a, for example, shows a net in which an unstable set of conditions exists. This causes the bar to fire, and the result of the firing is shown in Figure 2b. As a rule, the firing of a bar disables the circle conditions that cause the firing. A condition can be re-asserted, however, by connecting it to an output arc of the bar.

The arc leading to a bar may come from either an internal condition or an external condition, the latter being a condition whose state depends solely on the input values of the device. An external condition is represented in the net by a Boolean expression, and the condition is asserted if the expression has the value 1. In Figure 3, for example, the bar cannot fire unless $x = 0$ and $y = 1$. Because external conditions are controlled external to the device, the firing of a bar does not disable them and arcs may not lead to them. When the state of an external condition is to be indicated in a net, assertion of the condition is indicated by a marker to the left of the expression.

The arrival of a transition signal at a device input can be used to cause a change in the state of one or more internal conditions. This is represented in the net by a label (the name of the input) with outbound arcs leading to the circles representing the conditions that change. For example, Figure 4a shows the state of a net prior to the arrival of a transition signal on x, and Figure 4b shows the state after the arrival.

A device output that carries transition signals is represented by a label (the name of the output) with inbound arcs indicating the events that cause output signals. The event that causes the transition signal can be the firing of a bar or the arrival of a transition signal at a device input. In Figure 3, for example, a transition signal will occur on z if the bar fires or a signal arrives on w.

A device output that does not carry transition signals can be represented by a label (the output name) next to a circle. The condition represented by the circle is the output value 1. This representation is used only when it is convenient, that is, only when it does not complicate the net. Otherwise, the output is specified as a combinational function of the input variables and internal variables (conditions) of the device.

In addition to the representation rules already given, there are some additional rules that apply to the C-net as a whole.

1. The firing of a bar must disable one or more of the conditions that caused it to fire.
2. Bars with input arcs from the same circle must not be allowed to fire concurrently.
3. A bar may have only one input arc from a given circle and only one output arc to a given circle.
4. Multiple output arcs (from a bar, circle, or label) can be shown as individual arcs, as a single arc with branches, or as a combination of the preceding.

Rule 1 prohibits the continuous firing of a bar. Rule 2 prohibits a situation which would introduce races into the circuit realization. There are other situations in which concurrent events (bar firings, transition signal arrivals) are disallowed, and there are allowed situations which require special treatment during synthesis. The considerations are complex, however, and beyond the scope of this paper. Rule 3 eliminates unnecessary arcs and thereby makes the rules for circuit synthesis easier. Rule 4 is an analysis aid. By allowing branched arcs, the number of separate arcs and the number of arc cross-overs can be reduced, and this makes the net easier to understand.

There is one final rule: the C-net description of a device must indicate the initial state of each internal condition.

Examples

Figure 5 shows the diagram and description of the G element used by Chuang [5] in his synthesis procedure. The circle in the net is used to detect the arrival of signals on x. Each arrival causes assertion, and the subsequent firing of a bar causes non-assertion. The net shows that the arrival of a signal on x will cause a signal on output z_0 if d = 0 and will cause a signal on output z_1 if d = 1.

Figure 6 shows the diagram and description of a circuit used to detect the completion of concurrent control processes in the Washington University macromodules. As in the previous example, the circles in the net are used to detect the arrival of input signals. When signals have been received from each of the concurrent processes, the bar fires and causes a signal on output z. This firing also wipes out the memory of the input signals and returns the circuit to its initial state.

Figure 7 shows the description of a circuit that generates an output signal in response to every fourth input signal. The top circle is used to detect signals on x, and the lower four circles are used to count (modulo 4) the number of input signals detected. The dot in the lower left circle indicates that that circle is initially asserted. Assertion of that cell corresponds to zero signals received. Each signal received on x moves the marker one cell to the right, with the result that every fourth signal on x causes a signal on output z and returns the circuit to its initial state.

CIRCUIT REALIZATION

Normalization

A TL circuit description does not actually contain enough information to generate a unique level sequential circuit. A simple illustration of this is provided by the net in Figure 8 which specifies that each transition signal on x causes a transition signal on z. There are two distinct realizations for this net, namely $z = x$ and $z = \bar{x}$. The existance of more than one solution results from the fact that the net does not specify the initial values of the lines carrying transition signals. If we knew the initial values of z and x, there would be only one solution.

Our approach to this problem is to generate solutions which contain the initial values as variables. We do this by a normalization process that makes use of the following definitions:

Definition: For any variable x, the symbol $\underline{x}$ denotes the initial value of x.

Definition: For any variable x, the symbol X is a variable whose value is defined by the equation $X = \underline{x} \oplus x$. Variable X is called the normalized form of x.

The normalized form of a variable has two useful properties. First, its initial value is always 0 since $\underline{X} = \underline{x} \oplus \underline{x} = 0$. Second, it is functionally equivalent to the unnormalized form in applications where only changes in value are meaningful. This property derives from the fact that, for a given application, $\underline{x}$ is a constant. Thus $X = x$ or $X = \bar{x}$. In either case, a change in the value of X implies a change in the value of x, and vice versa. These two properties make it possible to synthesize a TL circuit in the following manner:

1. Normalize the circuit description by re-labeling the arcs carrying transition signals with the normalized form of the variable name.

2. Using the fact that the initial value of a normalized variable is 0, synthesize a solution for the circuit.

The solution thus formed is called a normalized solution. To get a circuit realization appropriate to a given application, one simply evaluates the normalized solution for that application. For example, the normalized solution for the net in Figure 8 is $Z = X$. Evaluation of this for an application where $\underline{z} = 0$ and $\underline{x} = 1$ gives $z = \bar{x}$.

Forming the Solution

We begin by assigning each circle a unique index i and each bar a unique index j. Circle i is treated as a variable c_i, with assertion of the circle corresponding to the condition $c_i = 1$. Bar j is treated as a variable y_j, with the firing of the bar corresponding to a change in the value of y_j. We then construct the function for each c_i and each y_j. Each c_i will be a combinational function of the device inputs and the y variables, and each y_j will be a single secondary variable. We then form the output functions of the device. These will be combinational functions of the device inputs and the y variables. The step by step procedure for the entire process is as follows.

1. Normalize the net description.

2. Assign each circle a unique index i and label the circle c_i. Assign each bar a unique index j and label the bar Y_j.

3. Generate the function for each c_i. This is done as follows: Let a_i be the EXCLUSIVE OR of the variables connected to the input arcs of the circle, and let b_i be the EXCLUSIVE OR of all bar variables connected to output arcs of the circle. Then

   $$c_i = a_i \oplus b_i \oplus \underline{c}_i$$

   where $\oplus$ denotes the EXCLUSIVE OR operator and where $\underline{c}_i = 1$ if and only if the circle is shown containing a marker.

4. Generate the excitation function for each Y_j. This is done as follows: Let f_j be the AND of the variables connected to the input arcs of the bar. If an arc comes from a circle, use the function generated in step 3 in forming the function f_j. Let s_j be the function f_j evaluated for $Y_j = 0$, and let r_j be the function f_j evaluated for $Y_j = 1$. Then the excitation function for Y_j is given by

   $$\overset{*}{Y}_j = s_j + Y_j \cdot \bar{r}_j$$

   where $\overset{*}{Y}_j$ represents variable Y_j at the next moment in time.

5. Generate the output function for each device output that carries transition signals. The output function is the EXCLUSIVE OR of all the variables connected to the output label by arcs.

6. Generate the output function for each remaining device output. Since the behavior of these outputs is already expressed as a combinational function of the circles and the device inputs, one need only replace the circle variables by the functions generated in step 3.

In the normalized solution constructed by the above process, secondary variable y_j occurs in normalized form. This is appropriate because the net specifies the value changes of y_j but not the values themselves. When the solution is evaluated for a given application, the initial value of y_j is not dictated by the application and can be chosen arbitrarily. One therefore chooses the value giving the simplest solution.

Examples

Circuit G The net for circuit G shown in Figure 9 has been labeled in accordance with steps 1 and 2 of the construction procedure. The calculations specified in the subsequent steps are as follows:

Step 3. $c_1 = X \oplus Y_0 \oplus Y_1$

Step 4. $f_0 = \bar{d} \cdot c_1 = \bar{d} \cdot (X \oplus Y_0 \oplus Y_1)$

$f_1 = d \cdot c_1 = d \cdot (X \oplus Y_0 \oplus Y_1)$

$\overset{*}{Y}_0 = \bar{d} \cdot (X \oplus Y_1) + Y_0 \cdot (d + (X \oplus Y_1))$

$\overset{*}{Y}_1 = d \cdot (X \oplus Y_0) + Y_1 \cdot (\bar{d} + (X \oplus Y_0))$

Step 5. $Z_0 = Y_0$

$Z_1 = Y_1$

Circuit R The net for circuit R shown in Figure 10 has been labeled in accordance with steps 1 and 2 of the construction procedure. The calculations specified in the subsequent steps are as follows:

Step 3. $c_1 = X_1 \oplus Y_1$

$c_2 = X_2 \oplus Y_1$

$c_3 = X_3 \oplus Y_1$

Step 4. $f_1 = c_1 \cdot c_2 \cdot c_3 = (X_1 \oplus Y_1) \cdot (X_2 \oplus Y_1) \cdot (X_3 \oplus Y_1)$

$\overset{*}{Y}_1 = X_1 \cdot X_2 \cdot X_3 + Y_1 \cdot (X_1 + X_2 + X_3)$

Step 5. $Z = Y_1$

COMMENTS AND OBSERVATIONS

Although the circuit realization procedure presented here is tightly wed to the net form of circuit description, this is actually a second marriage. The procedure was originally developed as the second part of a two part synthesis technique [9]. In the first part of that technique, a solution was generated which used special logic elements to detect and produce transition signals. In the second part, the special logic elements were eliminated by the realization procedure. The synthesis approach presented in this paper is a modification of that technique. It is structured around the Petri Net form of description for several reasons. First, this appears to be a convenient way of describing TL circuits, especially those used in concurrent control structures. Second, the net descriptions are sufficiently similar to the solutions formed in part I of [9] that the circuit equations can be derived directly from the description. Third, it provides a formalism which should be valuable in the treatment of races and in the development of techniques for minimizing the number of secondary variables.

The one-one correspondance between a bar in the net and a secondary variable in the realization has several important implications. First, there can be no secondary variable races if only one bar can fire at a time. Second, one approach for reducing the number of secondary variables is to develope a method for reducing the number of bars.

The synthesis technique does not depend on there being any inputs or outputs that carry transition signals. It can thus be applied to any level sequential circuit.

CONCLUSION

We have presented a procedure for synthesizing circuits whose behavioral descriptions are expressed by a form of Petri Net which we call a C-net. Although this technique was developed for the synthesis of transition-logic circuits, it applies to level sequential circuits in general. The procedure has the following characteristics:

1. The procedure works directly from the C-net description.
2. Designs are expressed in terms of ordinary combinational operators.
3. Solutions are generated in a matter of minutes.
4. Solutions are valid for any set of initial input and output values.
5. Numerous inputs and outputs do not present a problem.
6. The procedure is not restricted to circuits operated in fundamental mode.

REFERENCES

1. Clark, W.A., "Macromodule Computer Systems," 1967 Spring Joint Computer Conferences, AFIPS Proceedings Vol. 30, Thompson Book Company, Washington, D.C., (335-336), 1967.

2. Ornstein, S.M., Stucki, M.J., and Clark, W.A., "A Functional Description of Macromodules," 1967 Spring Joint Computer Conference, AFIPS Proceedings Vol. 30, Thompson Book Company, Washington, D.C., (337-355), 1967.

3. Stucki, M.J., Ornstein, S.M., and Clark, W.A., "Logical Design of Macromodules," 1967 Spring Joint Computer Conference, AFIPS Proceedings Vol. 30, Thompson Book Company, Washington, D.C., (357-364), 1967.

4. Huffman, D.A., "The Synthesis of Sequential Switching Circuits," Journal of the Franklin Institute, Vol. 257, No. 3, (161-190), No. 4, (275-303), March and April, 1954.

5. Chuang, Y.H., "Transition Logic Circuits and a Synthesis Method," IEEE Transactions on Computers, Vol. C-18, No. 2, (154-168), February, 1969.

6. Smith, J.R., and Roth, C.H., "Analysis and Synthesis of Asynchronous Sequential Networks Using Edge-Sensitive Flip-Flops," IEEE Transactions on Computers, Vol. C-20, No. 8, (847-855), August 1971.

7. Bredeson, J.G., and Hulina, P.T., "Synthesis of Multiple-Input Change Asynchronous Circuits Using Transition-Sensitive Flip-Flops," IEEE Transactions on Computers, Vol. C-22, No. 5, (524-531), May, 1973.

8. Holt, A.W., "Final Report of the Information System Theory Project," Technical Report RADC-TR-68-305, Rome Air Development Center, Griffiss Air Force Base, New York, 1968.

9. Stucki, M.J., "A Synthesis Approach for Transition Logic Circuits," Master's Thesis, Washington University, May, 1973.

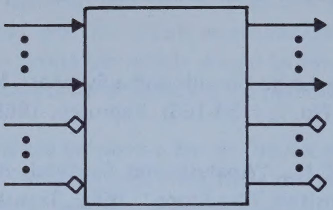


Figure 1

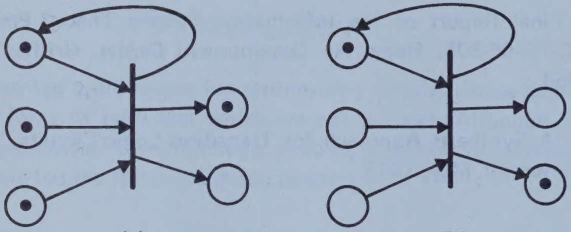


Figure 2

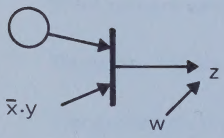


Figure 3

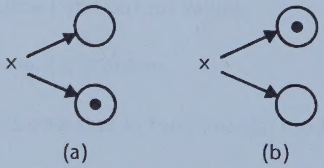


Figure 4

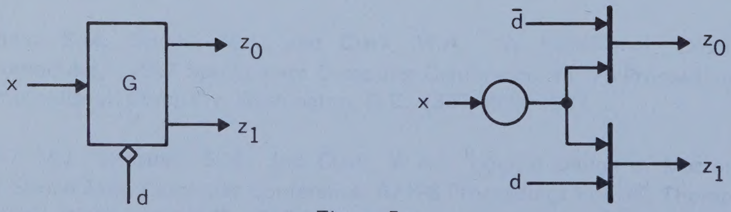

x
G
z_0
z_1
d
d̄
x
z_0
z_1
d

Figure 5

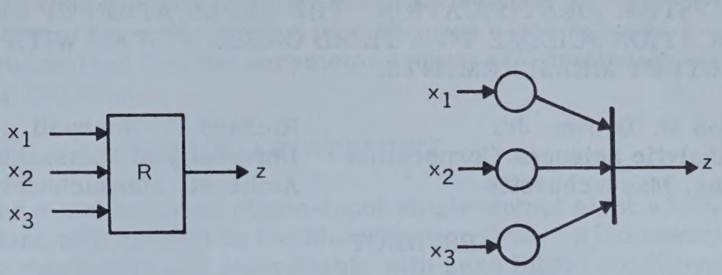

x1
x2
x3
R
z
x1
x2
x3
z

Figure 6

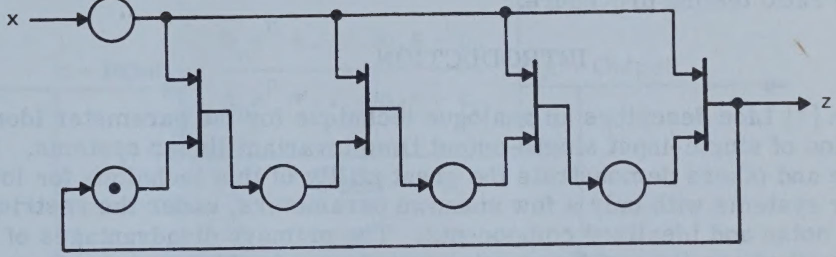


Figure 7

Figure 8

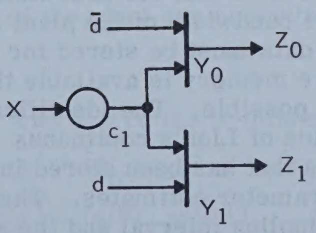

d̄
Z0
Y0
X
c1
Z1
d
Y1

Figure 9

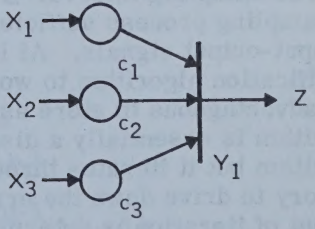

X1
c1
X2
Z
c2
Y1
X3
c3

Figure 10

LINEAR SYSTEM IDENTIFICATION: THE APPLICATION OF LION'S IDENTIFICATION SCHEME TO A THIRD ORDER SYSTEM WITH NOISY INPUT-OUTPUT MEASUREMENTS.

Charles M. Brown, Jr.
The Analytic Sciences Corporation
Reading, Massachusetts

Richard V. Monopoli
University of Massachusetts
Amherst, Massachusetts

ABSTRACT

A linear system identification technique developed by Lion [1] is adapted for use on a third-order system with six unknown parameters and noisy input-output measurements. A digital computer is employed so that rapid identification takes place with only two state variable filters. Bias in the parameter estimates is partially eliminated by a signal-to-noise ratio testing procedure.

INTRODUCTION

In [1] Lion describes an analogue technique for the parameter identification of single-input single-output time invariant linear systems. In [2] he and others demonstrate the great utility of this technique for low order systems with only a few unknown parameters, under the restriction of no noise and idealized components. The primary disadvantages of this method are its rapidly growing complexity for higher order systems with many unknown parameters and its production of biased parameter estimates in the presence of measurement noise. This paper assumes that a small digital computer is available to implement the technique and shows how the complexity of the method can be greatly reduced by making use of the computer's memory capability. With the computer a simple signal-to-noise ratio test can also be performed which partially reduces the bias in the parameter estimates due to an assumed level of measurement noise.

For rapid identification of an n^{th} order system with p unknown parameters Lion's method requires at least 2p n^{th} order state variable filters. The technique described here requires only 2 n^{th} order state variable filters. The states of each filter and its input are sampled every ΔT seconds and this data is stored by the digital computer as it comes in. The sampling interval ΔT must be chosen so that the bandwidth of the sampling process sufficiently covers the bandwidth of the plant and its input-output signals. At least p sets of data must be stored for the identification algorithm to work, but if more memory is available then it is advantageous to store as much data as possible. The identification algorithm is essentially a discretized version of Lion's continuous algorithm but it iterates through all the data that has been stored in memory to drive down the errors in the parameter estimates. The amount of iteration is determined by the sampling interval and the speed of the computer. If the parameter errors become so small that identification is essentially complete then the iteration stops. If noise is present the iteration continues until the error signal produced by the parameter errors becomes smaller than the assumed noise level in the error equation. When this occurs identification stops because it can no longer be guaranteed that the parameter errors will continue to decrease with further iteration.

PROBLEM STATEMENT

Consider a stable linear single-input single-output plant which is time invariant with respect to the identification time. Also assume that the plant is observable and controllable with zero initial conditions. The maximum possible order of the plant is n and it must have no poles at infinity. The only data available are noisy measurements of its input and output. Given these assumptions the plant may be represented by a transfer function as shown in Figure 1.

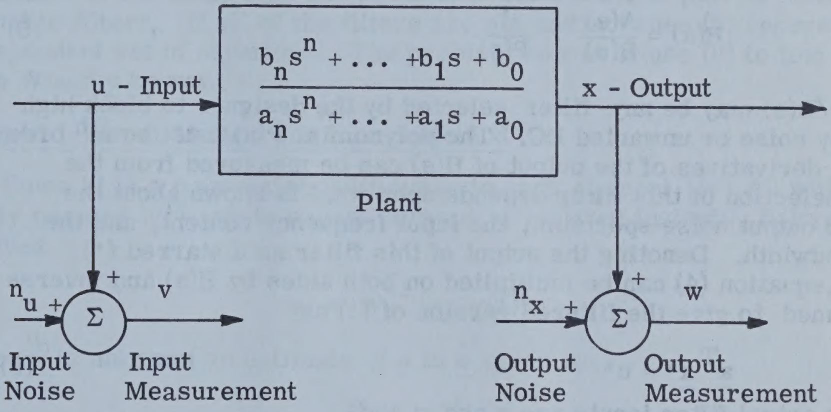


Figure 1

The problem is how to identify all of the a_i and b_i given v, w, and some rough knowledge of n_u and n_x. One of the plant coefficients must be assumed known to insure that there exists a unique set of coefficients to be identified, since multiplying each by a constant does not change the transfer function. The choice of which parameter to assume known depends upon what assumptions can be made about the plant. Here assume that $b_0 = 1$, then the DC gain cannot be zero. The differential equation of the plant can be written as

$$\underline{a}^T \underline{x} = u \tag{1}$$

where $\underline{a}$ and $\underline{x}$ are the p-dimensional vectors

$$\underline{a}^T = [a_n \; a_{n-1} \; .. \; a_1 \; a_0 \; -b_n \; -b_{n-1} \; ... \; -b_1] \tag{2}$$

$$\underline{x}^T = [x^{(n)} \; x^{(n-1)} \; ... \; \dot{x} \; x \; u^{(n)} \; u^{(n-1)} \; ... \; \dot{u}] \tag{3}$$

For zero initial conditions the Laplace Transform of (1) is

$$\underline{a}^T \underline{X}(s) = U(s) \quad ; \quad \underline{X}(s) = \mathcal{L}[\underline{x}] \tag{4}$$

Note that in general the right hand side of (1) is that signal whose coefficient was assumed known. The problem is to identify the p=2n+1 elements of $\underline{a}$ using the input-output measurements.

IDENTIFICATION

Equation (1) could be used to find $\underline{a}$ if all the signals in $\underline{x}$ were known, but since only noisy measurements of $\underline{x}$ and u are available this relation can not be used directly. Following Lion's development [1] we define state variable filters to be used on the plant input and output. These will give filtered estimates of the unavailable signals in (1). Define

$$H(s) = \frac{A(s)}{B(s)} \cdot \frac{1}{P(s)} \tag{5}$$

The A(s)/B(s) may be any filter selected by the designer to block high frequency noise or unwanted DC. The polynomial P(s) must be nth order so that n derivatives of the output of H(s) can be measured from the filter. Selection of this filter depends upon what is known about the input and output noise spectrum, the input frequency content, and the plant bandwidth. Denoting the output of this filter as a starred (*) variable,equation (4) can be multiplied on both sides by H(s) and inverse transformed to give the filtered version of (1) as

$$\underline{a}^T \underline{x}^* = u^* \tag{6}$$

Since the actual filter inputs are v and w and

$$v^* = u^* + n^* \qquad w^* = x^* + n_x^*$$

we can define:

$$\underline{w}^{*T} = \left[w^{(n)} \;\; w^{(n-1)} \;\ldots\; \dot{w} \;\; w \;\; v^{(n)} \;\; v^{(n-1)} \;\ldots\; \dot{v} \right]^*$$

$$\underline{n}^{*T} = \left[n_x^{(n)} \;\; n_x^{(n-1)} \;\ldots\; \dot{n}_x \;\; n_x \;\; n_u^{(n)} \;\; n_u^{(n-1)} \;\ldots\; \dot{n}_u \right]^*$$

$$\underline{w}^* = \underline{x}^* + \underline{n}^*$$

Equation (6) can be written as

$$\underline{a}^T (\underline{w}^* - \underline{n}^*) = v^* - n_u^*$$

or

$$v^* = \underline{w}^{*T} \underline{a} + \gamma^*(\underline{a})$$

with

$$\gamma^*(\underline{a}) = n^*_{\underline{u}} - \underline{n}^{*T}\underline{a}$$

In equation (7), $\gamma^*(\underline{a})$ is unknown colored noise, v^* and $\underline{w}^*$ are measured signals from the state variable filters and $\underline{a}$ is the unknown parameter vector to be identified. This equation is correct only if $\underline{a}$ is a constant and all system initial conditions have died out, but it will also be very nearly true if $\underline{a}$ varies very slowly with respect to the time constants of the plant and filters. This equation has p unknowns in $\underline{a}$ and unknown γ^*. To get p equations in p unknowns Lion defines p distinct sets of state variable filters so that (7) becomes a vector-matrix equation

$$\underline{v} = W\underline{a} + \underline{\gamma}(\underline{a}) \tag{9}$$

Here the (*) has been dropped and (9) represents the combination of p equations of the same type as (7) but each for a distinct pair of state variable filters. If all of the filters are not distinct then (9) represents a dependent set of equations. The problem now is to use (9) to find $\underline{a}$ with W and $\underline{v}$ known.

Method 1 Direct Method and Lion's Method

Since W is a p × p matrix with noise in each element $|W| \neq 0$ with probability one and W^{-1} exists except perhaps at isolated instants. Solving (9) for $\underline{a}$ gives

$$\underline{a} = W^{-1}[\underline{v} - \underline{\gamma}(\underline{a})] \tag{10}$$

Since $\underline{\gamma}$ is unknown an estimate of $\underline{a}$ is $\hat{\underline{a}}$ given by

$$\hat{\underline{a}} = W^{-1}\underline{v} \tag{11}$$

Define the estimation error as $\underline{\Delta} = \hat{\underline{a}} - \underline{a}$, then

$$\underline{\Delta} = W^{-1}\underline{\gamma} \tag{12}$$

When there is no noise and $|W| \neq 0$, we get $\underline{\Delta} = \underline{0}$, and $\hat{\underline{a}} = \underline{a}$. Equation (11) will give a good estimate of $\underline{a}$ only if W^{-1} exists and γ is very small. If the determinant of W is very small, the calculation of W^{-1} is very sensitive to noise in W and $\underline{\Delta}$ could be very large even in the low noise case.

Unfortunately (11) gives a biased estimate when the noise is not zero. To show this assume that x and u are both zero and only the measurement noise is driving the estimator. Then equation (12) becomes

$$\underline{\Delta} = \begin{bmatrix} \underline{n}^{*T} \\ \vdots \\ \vdots \end{bmatrix}^{-1} \begin{bmatrix} \underline{n}^{*T}\underline{a} \\ \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} \underline{n}^{*T} \\ \vdots \\ \vdots \end{bmatrix}^{-1} \begin{bmatrix} \underline{n}^{*T} \\ \vdots \\ \vdots \end{bmatrix} \underline{a} = \underline{a}$$

or

$$\hat{\underline{a}} = \underline{0}$$

Thus as the signal to noise ratio becomes very small the estimate is highly biased and independent of the true value of $\underline{a}$. Recalling equation (9), note the matrix W has components which are essentially filtered derivatives of the plant input and output measurements. These filtered derivatives may easily have very bad signal-to-noise ratios even when the original measurement noise is small. This causes the estimates of some of the elements of $\underline{a}$ to be strongly biased even in the low noise case. This problem becomes increasingly severe as the order of the plant increases and higher order filtered derivatives are required to estimate $\underline{a}$. In third and higher order systems measurement noise which seems insignificant often causes noticeable bias errors in some of the parameter estimates.

Lion's work assumes zero noise and he gives sufficient conditions on the input u so that the inverse of W will exist. Essentially u must contain enough frequency content to independently stimulate all of the modes of the plant. Lion does not solve (11) directly for $\underline{a}$ but uses a gradient approach. However, as the gains in this gradient method become large the method becomes essentially equivalent to solving for W^{-1} and calculating $\underline{a}$ directly. Noise causes the same bias to appear in Lion's method as in the direct method. The problems with these approaches can be summarized as follows: For more than second-order plants the estimates are strongly biased in zero mean noise. This bias is a complex function of the signal-to-noise ratios of the individual signals in the W matrix. These methods also require a bank of 2p n^{th} order filters for their implementation. The direct method requires the calculation of a $p \times p$ matrix inverse and Lion's method requires a gradient solution.

The following work partially overcomes these problems assuming that a small computer is available to process and store some of the measurements.

Method 2 A Practical Identifier

To reduce the hardware involved in the previous methods assume that only one pair of state variable filters is implemented. Then dropping the star notation equation (7) becomes

$$v = \underline{w}^T \underline{a} + \gamma(\underline{a}) \tag{13}$$

The measurements v and $\underline{w}$ are sampled every T seconds where T is small enough to pick up all frequencies in the bandwidth of the plant which are of interest. Then (13) can be written as follows, with $v_k = v(kT)$.

$$v_k = \underline{w}_k^T \underline{a} + \gamma_k(\underline{a}) \tag{14}$$

Now by storing the last N values of v_k and $\underline{w}_k$ there are N equations like (14) in memory available for estimating $\underline{a}$. This set of data can be used to improve upon the current estimate of $\underline{a}$. The amount of data stored depends upon the amount of memory available and the range of time in the past over which it can be assumed that $\underline{a}$ has remained essentially constant.

Between sampling instants the following algorithm is used to improve the estimate of $\underline{a}$.

$$\underline{\hat{a}}_{j+1} = \underline{\hat{a}}_j + \underline{w}_{j+1} (v_{j+1} - \underline{w}^T_{j+1} \underline{\hat{a}}_j)(s_{j+1}) \tag{15}$$

Here j is a data index and the algorithm iterates through the data set until the next sampling time when the new data is stored replacing the oldest data in the set. For the algorithm to work properly the gain parameter s_{j+1} must be chosen correctly in (15). The following development shows how to choose this gain so that $\underline{\hat{a}}_j$ converges towards the true value of $\underline{a}$. Begin by defining

$$\underline{\Delta}_j = \underline{a} - \underline{\hat{a}}_j \quad \text{and} \quad e_{j+1} = v_{j+1} - \underline{w}^T_{j+1} \underline{\hat{a}}_j \tag{16}$$

Then
$$e_{j+1} = \underline{w}^T_{j+1} \underline{\Delta}_j + \gamma_{j+1} \tag{17}$$

Combining (15) and (16) gives the following equation for the evolution of the parameter estimation error vector, $\underline{\Delta}_j$.

$$\underline{\Delta}_{j+1} = \underline{\Delta}_j - \underline{w}_{j+1} (\underline{w}^T_{j+1} \underline{\Delta}_j + \gamma_{j+1})(s_{j+1}) \tag{18}$$

Now select s_{j+1} such that the size of $\underline{\Delta}_{j+1}$ is less than or equal to $\underline{\Delta}_j$ for each iteration, that is choose s_{j+1} such that

$$\|\underline{\Delta}_{j+1}\| \le \|\Delta_j\| \tag{19}$$

Notice that this condition can always be satisfied by choosing $s_{j+1} = 0$.

By substituting (18) into (19) and using (17) the following criterion for choosing s_{j+1} can be derived assuming that s_{j+1} is positive.

$$0 < s_{j+1} \le \frac{2}{\|w_{j+1}\|^2} \left[1 - \frac{\gamma_{j+1}}{e_{j+1}}\right] \tag{20}$$

Thus if the right hand side of (20) is negative there is a contradiction since s_{j+1} was assumed positive. In this case s_{j+1} must be chosen equal to zero if condition (19) is not to be violated. Condition (20) is not very useful for selecting s_{j+1} because γ_{j+1} is an unknown noise quantity. However in practical applications it is usually possible to experimentally determine a bound for γ which is very rarely exceeded. Therefore assume that

$$|\gamma_j(\underline{a})| \le M \tag{21}$$

with very high probability. Given this assumption condition (20) can now be put into the following useful form.

$$0 \leq s_{j+1} \leq \frac{2}{\| \underline{w}_{j+1} \|^2} \left[1 - \frac{M}{|e_{j+1}|} \right] \qquad \text{if} \quad |e_{j+1}| \geq M \tag{22}$$

or

$$s_{j+1} = 0 \qquad \text{if} \quad |e_{j+1}| < M$$

Any value of s_{j+1} that obeys (22) insures that condition (19) is satisfied as long as (21) is not violated. Choosing s_{j+1} to be the midpoint of its allowed range has proved successful although there is no proof at the present time that this is the best value to choose. One possibility is to choose s_{j+1} randomly within its allowed range since the best value is not known. This has also worked well in simulations but it is not clear that this method is better than choosing the midpoint.

This technique has three principal advantages over those previously discussed. First, it requires only 2 nth order state variable filters. Second it requires that only equation (15) be implemented after the gain has been selected according to condition (22). Finally it guarantees that the estimate of $\underline{a}$ never diverges and given sufficient input signal u the estimate of $\underline{a}$ will always converge towards the correct value. Major disadvantages are the need for a digital computer and the bank of samplers required to sample all the states of the state variable filters. Another disadvantage is that the estimate will never completely converge to $\underline{a}$ because condition (22) sets the gain to zero when the error e_j gets below the noise cutoff level M.

EXPERIMENTAL WORK

This identification algorithm has been tested on several third order plants with six unknown parameters. In each case the transfer function of the plant was of the following form

$$G(s) = \frac{b_2 s^2 + b_1 s + 1}{a_3 s^3 + a_2 s^2 + a_1 s + a_0} \tag{23}$$

The technique works fairly well in low noise situations where the measurement noise amounts to less than one percent of the measured input and output signals. It is very sensitive to the frequency content of the input signal as are all identification techniques; it is not reasonable to expect any simple identifier to work if the input does not drive all of the plant's modes with sufficient strength so that the response of each of the modes can be detected in the plant output above the level of measurement noise.

The state variable filters must be chosen so that they have a bandwidth somewhat larger than that of the unknown plant. Good results were obtained in these experiments using third-order Butterworth filters. The sampling interval must be selected short enough so that all signals within the bandwidth of the plant are detected accurately, but not so short that the sampling picks up unwanted measurement noise outside of the plant bandwidth. The amount of memory required for data storage is not large but to get rapid identification under good conditions it is useful to store all the data received over the last two or three plant time constants.

The number of iterations of the identifying algorithm (15) between sampling intervals depends upon the speed of the computer in use, the length of the sampling interval, and the amount of data stored. Experience obtained for this third-order example with six unknown parameters indicates that ten iterations is usually sufficient to drive the error below its noise cutoff level M. The choice of M defined by condition (21) must be made experimentally for each particular application. M depends upon the magnitude and spectrum of the measurement noise, the range of possible parameter vectors $\underline{a}$, and the transfer function of the state variable filters. If M is chosen too small the algorithm can cause divergence of the parameter error vector and condition (19) may not hold for every iteration. If M is chosen too large, then the algorithm stops before the parameter errors have become as small as possible.

Given sufficient input frequency content and low noise level many experiments have shown that this technique can be made to work quite well. The actual speed of convergence can be made very fast, simply by using an input which drives the plant very hard so that the input and output signals are well above the measurement noise. The parameter estimates never completely converge to the true value of $\underline{a}$ because of the noise level cutoff feature of the filter, but with sufficient input the parameter errors become small enough that the transfer function of the estimated plant becomes nearly identical to that of the true plant.

CONCLUSIONS

A practical linear system identification technique has been developed that requires little hardware other than a small digital computer. It is computationally simple and can be applied to high order systems. With low measurement noise and a good plant input it works well. Its noise cutoff feature insures that the parameter estimates will not diverge if the input signal to the plant becomes small.

REFERENCES

1. Lion, P.M., "Rapid Identification of Linear and Nonlinear Systems," AIAA Journal, Vol. 5, No. 10, pp.1835-1842 (October 1967).

2. Hofmann, L.G., Lion, P.M., and Best, J.J., "Theoretical and Experimental Research on Parameter Tracking Systems," Systems Technology Inc., Hawthorne, Calif., TR-148-1 (October 1967).

ACKNOWLEDGEMENT

This research was supported in part by the National Aeronatics and Space Administration under grant NGL-22-010-018.

ON MINIMAL ERROR ENTROPY SYSTEM IDENTIFICATION

PAUL KALATA
Illinois Institute of Technology, Chicago, Illinois

ROLAND PRIEMER
University of Illinois, Chicago Circle, Chicago, Illinois

ABSTRACT

Information-theoretic concepts are used to stochastically identify a linear discrete time dynamic system based on noisy linear measurements of the system's state. It is interestingly noted that after various simplifying assumptions, the derived error entropy identification algorithm reduces to several well established identification techniques based on Kalman filtering, where statistical variances are replaced by respective entropy powers. Examples involving non-Gaussian statistics are used to compare several identification algorithms and illustrate the possible superiority of the error entropy identification procedure presented here.

I. INTRODUCTION

System identification or modeling is essential in the design and control of systems. Depending on the structure of the system, availability of input/output observations, and accessibility of an input control, there are many identification algorithms available [1] - [8]. This paper considers the problem of identifying the parameter $\underline{a}$ of the linear discrete time dynamic system described by

$$x(k) = a\,x(k-1) \tag{1}$$

where the noisy state observation procedure is modeled by

$$z(k) = h\,x(k) + v(k). \tag{2}$$

The system parameter $\underline{a}$ and the initial state x(o) are statistically independent random variables, and the observation noise v(k) is assumed to be a stationary zero mean white noise process, which is statistically independent of the system parameter $\underline{a}$ and the initial state x(0) for $k \geq 1$.

The identification problem consists of obtaining an estimate $\hat{a}(k)$ of the system parameter $\underline{a}$ given the sequence of observations $Z(k) = \{z(j), j = 1,2,\ldots,k\}$ such that the average of some reasonable cost function of the error

$$\tilde{a}(k) = a - \hat{a}(k) \tag{3}$$

is minimized. In this paper, information-theoretic concepts are used in devising a stochastic approximation approach to the identification problem.

The optimal mean square error (MSE) estimate of the parameter $\underline{a}$ given a sequence of observations Z(k) is the conditional mean, i.e.,

$$\hat{a}_{MSE}(k) = E(a/Z(k)),$$

where E(·) is the expectation operator. Since the relationship between the parameter $\underline{a}$ and the sequence of observations Z(k) is nonlinear and coupled with the fact that the observation noise sequence {v(k)} may be non-Gaussian, the conditional mean is intractable. Therefore, other approaches to the identification problem such as stochastic approximation methods are appealing.

Stochastic approximation is concerned with converging recursive schemes for fixed parameter estimation based on noisy observations and is closely related to optimal filtering theory. However, the stochastic approximation method is preferable in the identification problem because simpler calculations are involved. Compared to optimal filtering techniques, its main disadvantage is that the solution at each step is not optimal. However, if the stochastic approximation identification procedure converges, then the suboptimal approach is justified. Recently, Kalata and Priemer [9] used information-theoretic concepts in devising a stochastic approximation method to estimate the state of a nonlinear discrete time system, and in this paper, this approach is extended to the identification problem. For a more complete discussion of stochastic approximation theory, the reader can refer to Robbins and Monro [10] and Sakrison [11], and for its application to the identification problem to Ho and Lee [1], Saridis and Stein [4], and Sage and Melsa [8].

After presenting a few basic definitions in Sect. II, the use of the error entropy as the identification performance index is discussed. In Sect. III, a general error entropy stochastic approximation algorithm is developed, and after several simplifying assumptions it is shown that the algorithm reduces to an adaptive procedure similar to techniques presented elsewhere, which are based on the Kalman filtering procedure. In Sect. IV, examples involving non-Gaussian statistics are used to compare the error entropy identification algorithm with others based on Kalman filtering.

II. INFORMATION THEORY IN ESTIMATION

The term "information" is usually used to express the knowledge about an unknown contained in an observation. In the following, definitions of terms such as: average mutual information, entropy or average uncertainty, and average conditional uncertainty will be given. Usually, the concentration about the mean of a random variable $\underline{a}$ with probability density function p(a) is indicated by its variance

$$\sigma_a^2 = E((a-E(a))^2).$$

Another indicator of concentration is the entropy or average uncertainty of the ensemble defined by

$$H(a) = E(-\ln(p(a))),$$

where ln is the natural logarithmic function. In fact, for many ensembles there is a one-to-one relationship between ensemble entropy and variance. For example, for

$$\text{Gaussian:}\qquad H(a) = \tfrac{1}{2}\ln(2\pi e\sigma_a^2)$$

$$\text{Uniform:}\qquad H(a) = \tfrac{1}{2}\ln(12\sigma_a^2)$$

$$\text{Exponential:}\qquad H(a) = \tfrac{1}{2}\ln(e^2\sigma_a^2).$$

Given an observation z of the parameter $\underline{a}$ the average conditional uncertainty H(a/z) is similarly defined by

$$H(a/z) = E(-\ln(p(a/z))).$$

The average uncertainty about the parameter $\underline{a}$ resolved by the observation z is termed the average mutual information between $\underline{a}$ and z, and is defined by

$$I(a;z) = H(a) - H(a/z). \qquad (4)$$

Therefore, the term "information" is a measure of the decrease in uncertainty of the parameter $\underline{a}$ given the observation z. For an extensive discussion of Information Theory, the reader can refer to Gallager [12].

In the identification problem, consider the error entropy $H(\tilde{a}(k))$ as the cost criterion. If $\lim_{k\to\infty} H(\tilde{a}(k)) \to -\infty$, then the probability density function $p(\tilde{a}(k)) \to \delta(\tilde{a}(k))$, i.e. $\tilde{a}(k) \to o$ and identification is made with certainty. Using the error entropy criterion, Weidemann [13] established several filtering and feedback control principles which can be shown to apply to the identification problem. For example, without specifying the identification procedure, minimizing $H(\tilde{a}(k))$ is equivalent to minimizing $I(\tilde{a}(k); Z(k))$. A lower bound to the error entropy is the average conditional uncertainty of the parameter $\underline{a}$ given the observations $Z(k)$, i.e.

$$H(\tilde{a}(k)) \geq H(a/Z(k)). \tag{5}$$

This lower bound is achieved when $\tilde{a}(k)$ and $Z(k)$ are statistically independent which is analogous to the orthogonality condition in arriving at the minimal error variance filter for the normal estimation problem. Using information theoretic concepts, Kalata and Priemer [14] easily obtained a minimal error entropy recursive filter for the normal estimation problem (linear system, observation, and estimator and Gaussian statistics) which is identical to the well known Kalman filter.

III. ERROR ENTROPY IDENTIFICATION

Using (1) and (2) the state observation in recursive form becomes

$$z(k+1) = a\ z(k) + v(k+1) - a\ v(k)$$

for $k \geq 1$. Consider a sequence of parameter observations generated by successive ratios of state observations, i.e.,

$$\begin{aligned} z'(k+1) &= z(k+1)/z(k) \\ &= a + u(k+1), \end{aligned}$$

where $u(k+1) = (v(k+1) - a\ v(k))/z(k)$. Let the stochastic approximation approach in identifying the parameter $\underline{a}$ involve linear innovative improvements based on parameter observations, i.e.,

$$\hat{a}(k+1) = \hat{a}(k) + K(k+1)(z'(k+1) - \hat{a}(k)),$$

and the error becomes

$$\tilde{a}(k+1) = (1-K(k+1))\tilde{a}(k) - K(k+1)u(k+1).$$

The optimal filtering choice for the gain $K(k+1)$ is that value which minimizes the entropy of $\tilde{a}(k+1)$. But, in the stochastic approximation approach, only a sequence of gains $\{K(k+1)\}$ is to be determined such that the estimate $\hat{a}(k)$ converges to its true value in the error entropy sense.

In the continuous ensemble case, it is convenient to work with a derived quantity of the entropy called entropy power. Shannon [15] defines entropy power N as the power in a white noise process which is limited to the same frequency band as the original ensemble and having the same entropy H, i.e.,

$$N(\cdot) = \frac{1}{2\pi e} \exp(2H(\cdot)). \tag{6}$$

Since entropy power is a monotonic increasing function of the entropy, a minimization of entropy power simultaneously minimizes the entropy. Shannon [15] also showed that the entropy power N_3 of the sum of two independent ensembles with respective entropy powers N_1 and N_2 and respective variances σ_1^2 and σ_2^2 is bounded by

$$\sigma_1^2 + \sigma_2^2 \geq N_3 \geq N_1 + N_2. \tag{7}$$

Consider the selection of the gain sequence $\{K(k+1)\}$ by minimizing the error entropy power lower bound at each time step. Assume for the present that $\tilde{a}(k)$ and $u(k+1)$ are statistically independent, and therefore

$$N(\tilde{a}(k+1)) \geq N((1-K(k+1))\,\tilde{a}(k)) + N(-K(k+1)\,u(k+1)). \tag{8}$$

Using (6) and the identity

$$H(\xi) = H(\eta) - E(\ln|J(\eta/\xi)|), \tag{9}$$

where $J(\eta/\xi)$ is the Jacobian of the transformation, (8) becomes

$$N(\tilde{a}(k+1)) \geq (1 - K(k+1))^2\, N(\tilde{a}(k)) + K^2(k+1)\, N(u(k+1)). \tag{10}$$

The lower bound of $N(\tilde{a}(k+1))$ is minimal when

$$K(k+1) = \frac{N(\tilde{a}(k))}{N(\tilde{a}(k)) + N(u(k+1))}. \tag{11}$$

Using (4), (6), and (9), it can be shown that

$$N(u(k+1)) = N(v(k+1))\, \exp\,(2I(u(k+1);\, x(o),\, a,v(k)) - E(\ln(z^2(k)))).$$

Substituting the above into (10) and (11) and simplifying yields

$$N(\tilde{a}(k+1)) \geq \frac{N(\tilde{a}(k))N(v(k+1))\exp(2I(u(k+1);x(0),a,v(k)))}{N(v(k+1))\exp(2I(u(k+1);x(0),a,v(k)))+N(\tilde{a}(k))\exp(E(\ln(z^2(k))))} \tag{12}$$

$$K(k+1) = \frac{N(\tilde{a}(k))\exp(E(\ln(z^2(k))))}{N(v(k+1))\exp(2I(u(k+1);x(0),a,v(k)))+N(\tilde{a}(k))\exp(E(\ln(z^2(k))))}. \tag{13}$$

Depending on the statistical nature of the problem, the evaluation of $E(\ln(z^2(k)))$ and $I(u(k+1);\, x(o),\, a,v(k))$ may be difficult. However, in the case for which identification can be made with certainty, i.e., $|a| > 1$, several simplifying approximations will now be used to yield an adaptive error entropy stochastic approximation algorithm.

Assume that $|a| > 1$. For large k the system becomes unbounded, and therefore for purposes of computing the gain $K(k+1)$, the first simplifying approximation is

$$\exp(E(\ln(z^2(k)))) \cong z^2(k).$$

Since $x(0)$ is independent of $\underline{a}$ and $\{v(k)\}$, then it can be shown that

$$I(u(k+1);\, x(0),\, a,v(k)) \cong I(v(k+1) - a\, v(k);\, a,v(k)).$$

Using (4), (6), and (7), the above mutual information is bounded by

$$\tfrac{1}{2}\ln(1+\hat{a}^2(k)) \leq I(\cdot\ ;\ \cdot) \leq \tfrac{1}{2}\ln((1+\hat{a}^2(k))\ \sigma_v^2/N(v)).$$

It is obvious that the mutual information term in (12) and (13) is proportional to $\frac{1}{2}\ln(1+\hat{a}^2(k))$ and assuming the ratio $\sigma_v^2/N(v)$ is near unity, then

$$\exp(2I(u(k+1);\, x(o),\, a,v(k))) \cong 1 + \hat{a}^2(k)\ . \tag{14}$$

Substituting the approximations into (12) and (13) results in an adaptive identification algorithm

$$N(\tilde{a}(k+1)) \geq \frac{N(\tilde{a}(k))N(v(k+1))(1+\hat{a}^2(k))}{N(v(k+1))(1+\hat{a}^2(k)) + N(\tilde{a}(k))z^2(k)} \tag{15}$$

and

$$K(k+1) = \frac{N(\tilde{a}(k))z^2(k)}{N(v(k+1))(1+\hat{a}^2(k)) + N(\tilde{a}(k))z^2(k)}. \tag{16}$$

Except for the term $1 + \hat{a}^2(k)$, the procedure (16) for determining the gain sequence is similar to the adaptive algorithm of Ho and Lee [1], which is based on Kalman filtering techniques with statistical variances replaced by entropy powers. Furthermore, with Gaussian statistics, the ensemble variance and entropy power are identical, and the above identification procedure (except for the $1 + \hat{a}^2(k)$ term) is the Ho and Lee[1] procedure.

If the stochastic approximation approach in identifying the parameter $\underline{a}$ involves linear innovative improvements based on every other parameter observation, i.e.,

$$\hat{a}(k+2) = \hat{a}(k) + K(k+2)(z'(k+2) - \hat{a}(k)) \qquad (17)$$

for k = 0, 2, 4, ..., then, following the same above procedures and assumptions, the adaptive identification algorithm becomes

$$N(\tilde{a}(k+2)) \geq \frac{N(\tilde{a}(k))N(v(k+2))(1+\hat{a}^2(k))}{N(v(k+2))(1+\hat{a}^2(k)) + N(\tilde{a}(k))z^2(k+1)} \qquad (18)$$

and

$$K(k+2) = \frac{N(\tilde{a}(k))z^2(k+1)}{N(v(k+2))(1+\hat{a}^2(k)) + N(\tilde{a}(k))z^2(k+1)} . \qquad (19)$$

Except for the term $1 + \hat{a}^2(k)$ in eqs. (18) and (19), the adaptive identification algorithm is similar to the algorithm of Saridis and Stein [4] (which includes an unbiasing term) with statistical variances replaced by entropy powers.

Since the error entropy adaptive identification algorithms were developed under the assumption that $|a| > 1$, then it is apparent that they would be less efficient for the case in which $|a| < 1$. In fact, it is doubtful whether there exists any algorithm which can identify the parameter $\underline{a}$ with certainty when $|a| < 1$. If identification is to be made with certainty, then $H(\tilde{a}(k)) \to -\infty$ and for its lower bound, as given by (5), $H(a|Z(k)) \to -\infty$. Using (4), the lower bound, which is independent of any identification procedure, becomes

$$H(a|Z(k)) = H(a) - I(a;Z(k)). \qquad (20)$$

Using the identity

$$I(\xi;\eta,\gamma) = I(\xi;\eta) + I(\xi;\gamma|\eta),$$

then (20) becomes

$$H(a|Z(k)) = H(a) - \sum_{j=1}^{k} I(a;z(j)|Z(j-1)). \qquad (21)$$

Assuming that the parameter $\underline{a}$ has some finite uncertainty, i.e., $|H(a)| < \infty$, then identification with certainty is dependent on the divergence of the conditional mutual information series sum in (21). The series inherently satisfies the necessary convergent condition that the k^{th} term must vanish as $k \to \infty$. For large k, $z(k) \to v(k)$, and since $\underline{a}$ is independent of the noise process $\{v(k)\}$, then

$$I(a;z(k)|Z(k-1)) \to 0.$$

The mutual information $I(a;Z(k))$ can be bounded from above by considering further the summation in (21). Each conditional mutual information term in (21) has a maximum value of the order

$$\max\ I(a;z(j)|Z(j-1)) \cong \tfrac{1}{2}\ln(\sigma_v^{-2}\,\sigma^2_{z(j)|Z(j-1)}) \qquad (22)$$

where equality exists when the random variable $z(j)|Z(j-1)$ is normally distributed. For a normal random variable it can be shown that

$$\sigma^2_{z(j)|Z(j-1)} \leq \sigma^2_{z(j)} = \sigma^2_{x(0)}\sigma_a^{2j}\,h^2 + \sigma_v^2 .$$

Substituting the above inequality into (22) yields

$$I(a;z(j)|Z(j-1)) \leq \tfrac{1}{2}\ln(1 + \sigma_v^{-2}\sigma^2_{x(0)}\sigma_a^{2j}\,h^2) .$$

For any real number $\alpha > 0$, $\ln(1+\alpha) < \alpha$, and therefore

$$I(a;z(j)|Z(j-1)) < \tfrac{1}{2}\,\sigma_v^{-2}\,\sigma^2_{x(0)}\,\sigma^2_{a^j}\,h^2 . \tag{23}$$

Inequality (23) is useful for showing the convergence of the series in (21).

For example, let the parameter $\underline{a}$ be uniformly distributed over $[0, 1 - \varepsilon]$ where $\varepsilon > 0$ prohibits $\underline{a}$ from achieving unity. The variance of a^j becomes

$$\sigma^2_{a^j} = \frac{j^2(1-\varepsilon)^{2j}}{2j^3 + 5j^2 + 4j + 1} < \frac{(1-\varepsilon)^{2j}}{2j} .$$

Substituting the above into (23) yields

$$I(a;z(j)|Z(j-1)) < \frac{\sigma^2_{x(0)}(1-\varepsilon)^{2j}h^2}{4j\,\sigma_v^2} ,$$

and therefore (21) becomes

$$H(a|Z(k)) > H(a) - \frac{\sigma^2_{x(0)}\,h^2}{4\,\sigma_v^2}\sum_{j=1}^{k}\frac{(1-\varepsilon)^{2j}}{j} .$$

It can be shown that the infinite series $\{\frac{(1-\varepsilon)^{2j}}{j}\}$ is convergent, and assuming that $H(a)$, σ_v^2, and $\sigma^2_{x(0)}$ are finite, then

$$\lim_{k \to \infty} H(a|Z(k)) > -\infty .$$

Since $H(\tilde{a}(k)) \geq H(a|Z(k))$, the identification of the parameter $\underline{a}$ cannot be made with certainty when $|a| < 1$.

IV. EXAMPLES

As it is with many estimation procedures, the efficiency of the error entropy adaptive identification algorithms is dependent on the signal-to-noise ratio (S/N) of the measurement procedure. The system (1-2), h=1., was simulated with the measurement noise process $\{v(k)\}$ generated by a uniformly distibuted random number generator with zero mean and unity variance and entropy of 1.2425. Using the error entropy adaptive identification algorithm described by (15) and (16) (EE I), Fig. 1 illustrates the convergence nature of the algorithm for several combinations of x(o) and $\underline{a}$. With a constant S/N of unity (x(o) = 1.0, $\underline{a}$ = 1.0) the estimate of $\underline{a}$ is approximately 50% of its true value and is converging at a slow rate after 50 observations. For a constant S/N of 100.0 (x(o) = 10.0, $\underline{a}$ = 1.0), the estimate of $\underline{a}$ is approximately 95% of its true value and is also converging slowly. With an exponentially increasing S/N (x(o) = 1.0, $\underline{a}$ = 1.1), then after 50 observations, the estimate of $\underline{a}$ is approximately 99% of its true value and is converging rapidly.

A Monte Carlo simulation (500 trials) of (1) and (2) was made by simulating the parameter $\underline{a}$ by an exponentially distributed random number generator with unity mean, variance, and entropy, using x(o) = 1, and using the measurement noise process as described above. The performance of the error entropy adaptive identification algorithms EE I and that described by (18) and (19) (EE II) is shown in Fig. 2 (average error) and Fig. 3 (error variance). For a comparison, the results of the adaptive identification algorithm of Ho and Lee [1] also appears in these Figures. The algorithms of Saridis and Stein [4], Budin [7], and Mehra [5] were also tested under the same conditions. Table I lists the final sample identification average errors and error variances for the six algorithms.

As a second example, the parameter a was simulated by a normally distributed random number generator with unity mean, variance and entropy power with the remaining specifications of the previous example. Figs. 4 and 5 are again the respective identification average error and error variance of the Monte Carlo simulation. Although the statistical nature of the problem becomes normal, the difference between the Ho and Lee [1] algorithm and that of the EE I algorithm is still significant.

V. CONCLUSION

This paper considered the involvement of several information-theoretic concepts in the stochastic approximation approach to system identification. The entropy power of the identification error was used as a performance index to obtain a sequence of gains for the stochastic approximation procedure. After some simplifying assumptions and approximations, the developed general algorithm reduced to an adaptive identification algorithm similar to an algorithm based on Kalman filtering theory.

A conditional entropy, which is the identification error entropy lower bound, was used in devising a procedure to determine when identification of the parameter a, $|a| < 1$, cannot be made with certainty. Although only a uniform distribution was considered as an example, for many other distributions identification cannot be made with certainty.

For particular cases, Monte Carlo simulations involving non-Gaussian statistics were made to test the error entropy adaptive identification algorithms as well as several other identification procedures. The resulting error statistics show that the error entropy identification algorithms compare favorably to the other identification algorithms used in the comparison.

REFERENCES

[1] Y. C. Ho and R. C. K. Lee, "Identification of Linear Dynamic Systems," J. Inform. Contr., vol. 8, Feb., 1965, pp. 93-110.

[2] K. Y. Wong and E. Polak, "Identification of Linear Discrete Time Systems Using the Instrumental Variable Method," IEEE Trans., vol. AC-12, Dec., 1967, pp. 707-718.

[3] D. J. Sakrison, "The Use of Stochastic Approximation to Solve The System Identification Problem," IEEE Trans., vol. AC-12, Oct., 1967, pp. 563-567.

[4] G. N. Saridis and G. Stein, "Stochastic Approximation Algorithms for Linear Discrete-Time System Identification," IEEE Trans., vol. AC-13, Oct., 1968, pp. 515-523.

[5] R. K. Mehra, "On-Line Identification of Linear Dynamic Systems with Applications to Kalman Filtering," IEEE Trans., vol. AC-16, Feb., 1971, pp. 12-21.

[6] J. B. Farison, R. E. Graham and R. C. Shelton, Jr., "Identification and Control of Linear Discrete Systems," IEEE Trans., vol. AC-12, Aug., 1967, pp. 438-442.

[7] M. A. Budin, "A New Approach to System Identification and State Estimation," IEEE Trans., vol. SMC-2, July, 1972, pp. 396-401.

[8] A. P. Sage and J. L. Melsa, System Identification, New York: Academic Press, 1971.

[9] P. Kalata and R. Priemer, "On Minimal Error Entropy Stochastic Approximation," submitted for publication.

[10] H. Robbins and S. Monro, "A Stochastic Approximation Method," Ann. of Math. Statist., vol. 22, 1951, pp. 400-407.

[11] D. J. Sakrison, "Stochastic Approximation: A Recursive Method for Solving Regression Problems," Adv. Comm. Systems, A. V. Balakrishnan(ed.), vol. 2, 1966, pp. 51-106.

[12] R. G. Gallager, Information Theory and Reliable Communication, New York: Wiley, 1968.
[13] H. L. Weidemann, "Entropy Analysis of Feedback Control Systems," Adv. Contr. Systems, C. T. Leondes(ed.), vol. 7, 1969, pp. 225-255.
[14] P. Kalata and R. Priemer, "On the Estimation Problem as Viewed from Information Theory," Proc. 10th Annu. Allerton Conf. Circuit and System Theory, 1972, pp. 99-100.
[15] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, U. of Ill. Press, Urbana, Illinois, 1949.

Table I. Final Error Statistics of the First Example

Algorithm	$E(\tilde{a}(50))$	$VAR(\tilde{a}(50))$
Ho and Lee[1]	.2244	.0753
EE I	.1711	.0623
EE II	.1302	.0677
Saridis and Stein[4]	-.9989	7.2352
Budin[7]	-1.2801	811.25
Mehra[5]	.5977	138.64

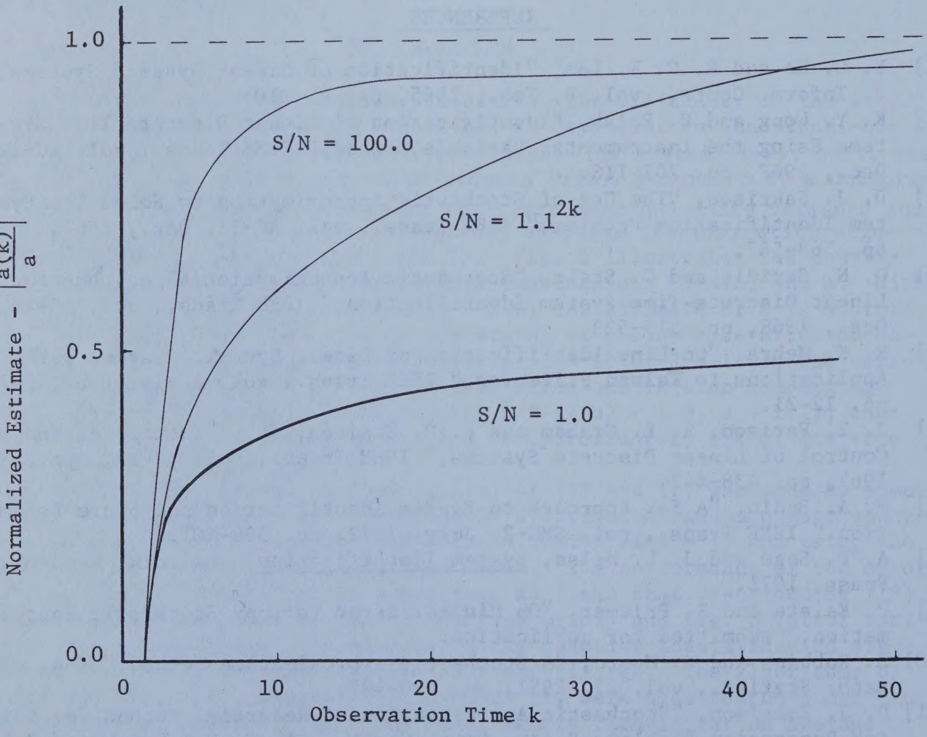


Fig. 1. Efficiency Of EE I.

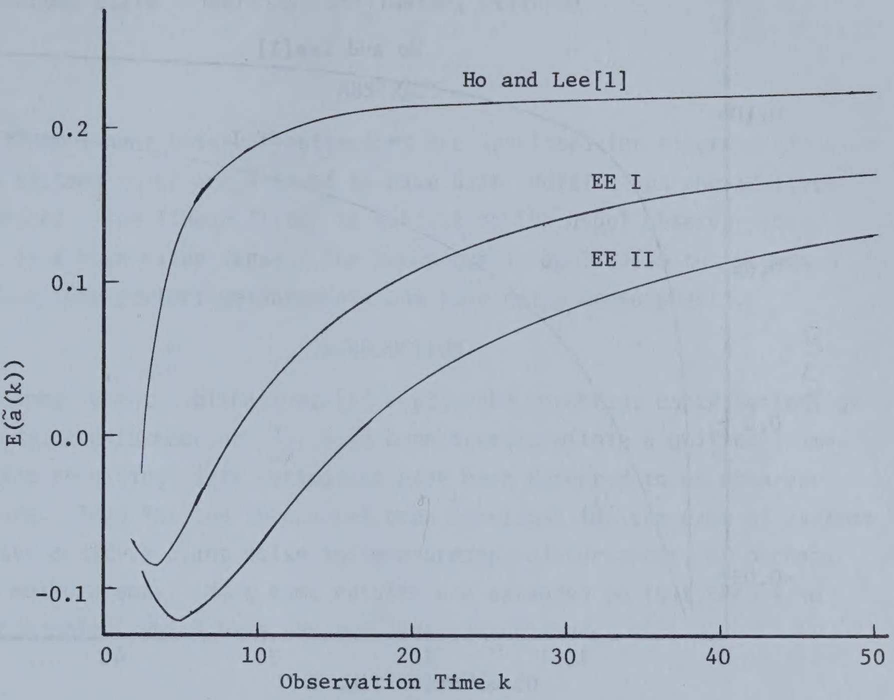


Fig. 2. Average Error For The First Example.

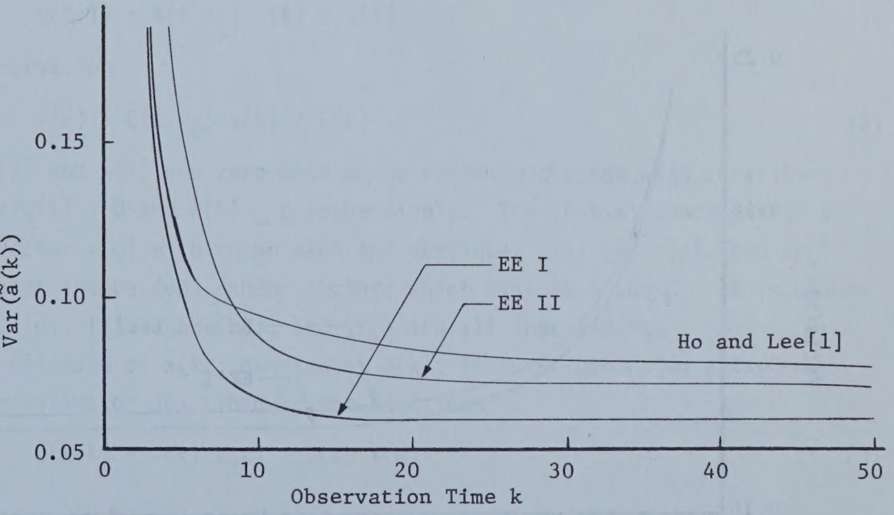


Fig. 3. Error Variance For The First Example.

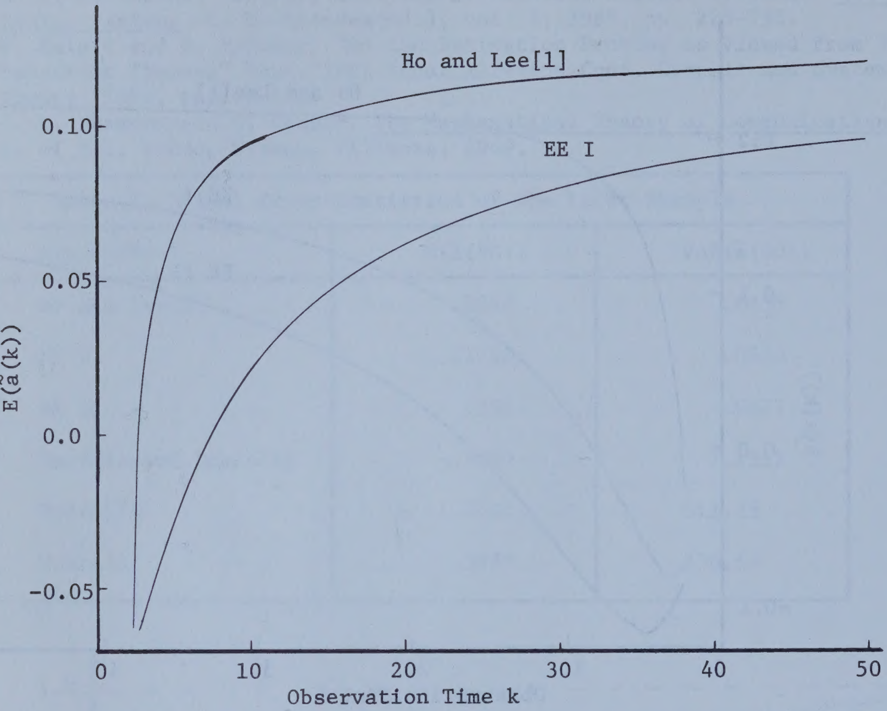


Fig. 4. Average Error For The Second Example.

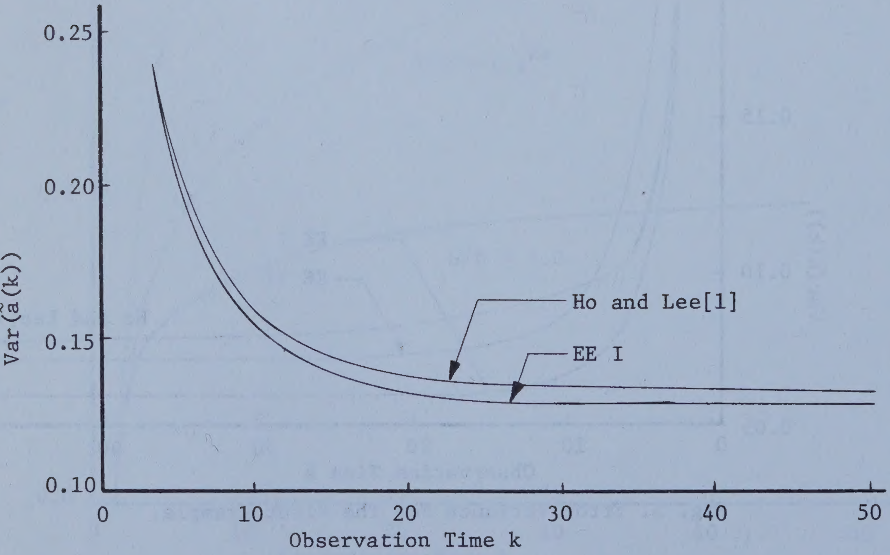


Fig. 5. Error Variance For The Second Example.

OBSERVER-ESTIMATORS FOR DISCRETE RANDOM PARAMETER SYSTEMS

CRAIG S. SIMS
Oklahoma State University, Stillwater, Oklahoma

ABSTRACT

Optimum linear observer-estimators are developed for discrete time-varying systems which are allowed to have both additive and non-additive disturbances. The linear filter is subject to the usual observer constraint in a mean value sense. The technique is applicable to systems which have some perfect measurements and some noisy measurements.

INTRODUCTION

In some recent publications [1] - [2], the important contributions of Kalman [3], and Luenberger [4], have been treated within a unified framework. The resulting state estimators have been referred to as observer-estimators. Thus far the theory has been developed for the case of systems which have additive plant noise and measurement disturbances, or perhaps perfect measurements. Here some results are extended so that they also apply to problems which have non-additive disturbances.

PROBLEM STATEMENT

We consider a linear dynamical system of the form

$$x(k+1) = A(k,r_k)\, x(k) + \xi(k) \tag{1}$$

with observation

$$Y(k) = C(k,r_k)\, x(k) + \eta(k) \tag{2}$$

where $\xi(k)$ and $\eta(k)$ are zero-mean white random processes with covariance matrices $Q(k) \geq 0$ and $R(k) \geq 0$ respectively. The initial condition is a random vector $x(o)$ with known mean and variance, $\mu(o)$ and $\Sigma(o)$, and $\{r_k\}$ is a set of independent random vectors which disturb A and C. It is assumed that $x(o)$, $\{\xi(k)\}$, $\{\eta(k)\}$, and $\{r_k\}$ are all independent.

An estimate of $x(k)$, denoted as $W(k)$, is to be generated using an observer-estimator described by the equations

$$Z(k+1) = F(k)\, Z(k) + D(k)\, Y(k) \tag{3}$$

and

$$W(k) = P(k)\, Z(k) + V(k)\, Y(k) \tag{4}$$

where P(k) and V(k) are assumed to exist such that

$$P(k)\ T(k) + V(k)\ E\{C(k,r_k)\} = I \tag{5}$$

which is in effect an observer constraint. If $x \varepsilon R^n$, and $y \varepsilon R^m$, then Z is of dimension $S \geq n-m$. The observer-estimator is to have the properties that

$$E\{Z(k)\} = T(k)\ E\{x(k)\} \tag{6}$$

$$E\{W(k)\} = E\{x(k)\} \tag{7}$$

and a quadratic error criterion

$$J = E\{[W(k+1) - x(k+1)]^T[W(k+1) - x(k+1)]\}\} \tag{8}$$

is to be minimized. Equations (5) and (6) are requirements such that the observer-estimator is equivalent to an ordinary observer in a mean value sense. Equation (7) is simply a requirement that the estimate be unbiased. In the development presented here, the performance measure, (8), is minimized using a single stage optimization procedure.

UNBIASED REQUIREMENTS

In this section it is shown that Eqs. (6) and (7) can be satisfied provided that F(k) and D(k) are selected as

$$F(k) = T(k+1)\ E\{A(k,r_k)\}\ P(k) \tag{9}$$

$$D(k) = T(k+1)\ E\{A(k,r_k)\}\ V(k) \tag{10}$$

It is convenient to consider the difference term, $Z(k+1) - T(k+1)\ x(k+1)$. Substitution from Eqs. (9) and (10) in Eq. (3) results in the expression

$$\begin{aligned} Z(k+1) - T(k+1)\ x(k+1) &= T(k+1)\ \overline{A}(k)\ P(k)\ Z(k) \\ &+T(k+1)\ \overline{A}(k)\ V(k)\ Y(k) - T(k+1)\ x(k+1) \end{aligned} \tag{11}$$

where we have adopted the convenient notation $\overline{A}(k) = E\{A(k,r_k)\}$. Further substitution, from Eqs. (1) and (2) provides the difference equation,

$$\begin{aligned} &Z(k+1) - T(k+1)\ x(k+1) = T(k+1)\ \overline{A}(k)\ [P(k)\ Z(k) \\ &+V(k)\ C(k,r_k)\ x(k) + V(k)\ \eta(k)] - T(k+1)\ [A(k,r_k)\ x(k) + \xi(k)] \end{aligned} \tag{12}$$

Upon taking expectations, we obtain

$$\begin{aligned} E\{Z(k+1) - T(k+1)\ x(k+1)\} &= T(k+1)\ \overline{A}(k)\ [P(k)\ E\{Z(k)\} \\ &+ (V(k)\ \overline{C}(k) - I)\ E\{x(k)\}] \end{aligned} \tag{13}$$

where $\overline{C}(k) = E\{C(k,r_k)\}$. Hence, using Eq. (5), it is clear that

$$E\{Z(k+1)\} - T(k+1)\ E\{x(k+1)\} = T(k+1)\ \overline{A}(k)\ P(k)[E\{Z(k)\} - T(k)\ E\{x(k)\}] \tag{14}$$

From Eq. (14), it can then be seen that Eq. (6) will be satisfied provided that the initial condition $Z(o) = T(o)\mu(o)$.

It is not difficult to show that the estimate is unbiased. Taking the expectation of Eq. (4), there results

$$E\{W(k)\} = P(k)\ E\{Z(k)\} + V(k)\ \overline{C}(k)\ E\{x(k)\} \tag{15}$$

From Eq. (6) it then follows that

$$E\{W(k)\} = [P(k)\ T(k) + V(k)\ \overline{C}(k)]\ E\{x(k)\} \tag{16}$$

and Eq. (7) is satisfied whenever the observer constraint, Eq. (5), is satisfied. The selection of F(k) and D(k) according to Eqs. (9) and (10) is then appropriate in that it leads to the design objectives expressed by Eqs. (6) and (7). Thus in the truly stochastic case the observer-estimator provides an unbiased estimate while in the degenerate deterministic case, we have an ordinary observer.

PERFORMANCE OPTIMIZATION

In this section the objective is to choose a matrix V in such a way as to minimize the performance measure indicated by Eq. (8). To accomplish this it is convenient to consider the dynamics for the error, $e(k) \underline{\Delta}\ x(k) - W(k)$. By substitution from Eqs. (1), (2), and (4) we have

$$\begin{aligned} e(k+1) = {} & [I - V(k+1)\ C(k+1,r_{k+1})][A(k,r_k)\ x(k) + \xi(k) \\ & -P(k+1)\ Z(k+1) - V(k+1)\ \eta(k+1) \end{aligned} \tag{17}$$

Further manipulation yields

$$\begin{aligned} e(k+1) = {} & [I - V(k+1)\ C(k+1,r_{k+1})][A(k,r_k)\ e(k) + \xi(k)] \\ & - V(k+1)\ \eta(k+1) + [P(k+1\ T(k+1)\ \tilde{A}(k) - V(k+1)\ \tilde{C}(k+1)\ A(k,r_k)]\ W(k) \end{aligned} \tag{18}$$

where

$$\tilde{A}(k)\ \underline{\Delta}\ A(k,r_k) - \overline{A}(k)$$

$$\tilde{C}(k)\ \underline{\Delta}\ C(k,r_k) - \overline{C}(k)$$

Substituting for W(k) and simplifying we obtain

$$\begin{aligned} e(k+1) = {} & [I - V(k+1)\ \overline{C}(k+1)][\overline{A}(k)\ e(k) + \xi(k)] - V(k+1)\eta(k+1) \\ & + [\tilde{A}(k) - V(k+1)\ \tilde{\Gamma}(k+1)]\ x(k) \end{aligned} \tag{19}$$

where the new notation

$$\tilde{\Gamma}(k+1) \underline{\Delta} C(k+1,r_{k+1}) A(k,r_k) - \bar{C}(k+1) \bar{A}(k)$$

has been introduced.

The performance measure (8) can be written as the trace of the error covariance matrix at stage k+1.

$$J = \mathrm{tr}[E\{e(k+1) e(k+1)^T\}] = \mathrm{tr} \{\Sigma(k+1)\} \tag{20}$$

By multiplying each side of Eq. (19) by its transpose and taking expected values the propogation mechanism of the error covariance matrix is obtained.

$$\begin{aligned}\Sigma(k+1) = {} & [I - V(k+1) \bar{C}(k+1)] \psi(k) [I - V(k+1) \bar{C}(k+1)]^T + E\{[\tilde{A}(k) \\ & - V(k+1) \tilde{\Gamma}(k+1)] M(k) [\tilde{A}(k) - V(k+1) \tilde{\Gamma}(k+1)]^T\} \\ & + V(k+1) R(k+1) V^T(k+1)\end{aligned} \tag{21}$$

where $M(k)$ and $\psi(k)$ are defined as

$$M(k) = E\{x(k) x(k)^T\}$$

and

$$\psi(k) = \bar{A}(k) \Sigma(k) \bar{A}(k)^T + Q(k)$$

Note that $M(k)$ can be precomputed from the matrix difference equation

$$M(k+1) = E\{A(k,r_k) M(k) A^T(k,r_k)\} + Q(k) \tag{22}$$

with the initial condition

$$M(o) = \Sigma(o) + \mu(o) \mu^T(o)$$

Hence it can be regarded as known. The problem is to choose $V(k+1)$ to minimize J subject to the constraint equation (21). The approach used here is simply to substitute Eq. (21) in Eq. (20) and set $\frac{\partial J}{\partial V(k+1)} = 0$, as a necessary condition for optimality. Note that the problem could also be cast as a final value control problem and solved using the discrete matrix minimum principle as in [5]. Using the indicated single stage optimization procedure results in the equation which $V(k+1)$ must satisfy

$$\begin{aligned}\psi(k) \bar{C}^T(k+1) + E\{\tilde{A}(k) M(k) \tilde{\Gamma}^T(k+1)\} = {} & V(k+1) [\bar{C}(k+1) \psi(k)\bar{C}^T(k+1) \\ & + R(k+1) + E\{\tilde{\Gamma}(k+1) M(k) \tilde{\Gamma}^T(k+1)\}]\end{aligned} \tag{23}$$

or

$$\Lambda^T(k+1) = V(k+1) \Xi(k+1) \tag{24}$$

where the definitions of Λ and Ξ are obvious from Eq. (23). If Eq. (24)

has a solution, which it does if

$$\text{rank } \Xi(k+1) = \text{rank } [\Xi(k+1) \mid \Lambda(k+1)]$$

then the general expression for the V(k+1) which solves Eq. (24) is

$$V^T(k+1) = \Xi^\dagger(k+1)\, \Lambda\, (k+1) + [I - \Xi^\dagger\, (k+1)\, \Xi\, (k+1)]\, G \tag{25}$$

In the above equation † indicates the pseudo-inverse and G is an arbitrary mxn matrix. If Eq. (25) has a unique solution, then we have established an optimum sequence of observer gains $\{V(k)\}$ which can be computed by recursively solving Eqs. (21) and (22) with the optimum choice of V used at each stage. It is clear that in general the sequence $\{V(k)\}$ is not unique.

EXAMPLE PROBLEM

The example considered here represents a simplified version of a discretized inertial system. The discrete equations describing the dynamics of the system are

$$x(k+1) = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \xi(k) \tag{26}$$

where Δ is .02 and the observation model is

$$Z(k+1) = (1+r_{k+1})\, x_1(k+1) + \eta(k+1) \tag{27}$$

The noise $\{\xi(k)\}$ is a white sequence with covariance parameter 1 and $\{\eta(k)\}$ is a white noise sequence with covariance parameter .1, while $\{r_k\}$ is a set of independent random variables such that

$$r_k = \alpha \text{ with probability } 1/2$$

$$r_k = -\alpha \text{ with probability } 1/2$$

The initial variance of the state vector is assumed to be the identity matrix, and the initial mean value of each state is zero. The equations described herein were applied with $\alpha = 1$. After 10 stages the variance of error in estimating x_1 is .151 and the variance of error in estimating x_2 is .529. At the same stage if a Kalman filter is used (designed on the basis that r_k is always at its mean value of zero) the error variance is 33.8 for x_1 and 219.1 for x_2. This dramatic difference is of course an indication that one cannot simply neglect multiplicative noise. At later stages, the difference continues to grow.

The observer-estimator for this example is second order, which it must be if we are allowed to optimize with respect to $\{V(k)\}$. Note that it does not provide a truly optimal estimate and performance deteriorates at later stages. The observer-estimator gives the optimum estimate subject to its structural constraint, which is certainly not the unconstrained optimum in general.

CONCLUSION

We have considered the design of optimum linear observer-estimators for systems with both additive and non-additive disturbances. The estimate generated by the proposed algorithm is unbiased. An equation for the propogation of the error covariance matrix has been derived and a one stage optimization procedure used to optimally select V(k), such that a quadratic error criterion is minimized. The result is not unique in general, and more complicated than the optimum gain obtained when only additive disturbances are allowed.

The results are more general than those obtained in [5] in the sense that a lower order filter may sometimes be used. The estimator is equivalent to that obtained in [5] if R(k) > 0, and equivalent to that obtained in [1] when only additive noise is present. The estimate will not be optimum in general, even when all noise terms and initial conditions are gaussian, since the state and observation may be non-gaussian due to the non-additive disturbances. Hence a conditional mean estimate is not generated by the linear structure.

Although the algorithms presented here are reasonable from a viewpoint of computational complexity, the appropriateness of a "best linear" type of design must be judged according to the particular problem considered.

REFERENCES

[1] Tse, E., and M. Athans, "Optimal Minimal-Order Observer-Estimators for Discrete Linear Time Varying Systems," IEEE Trans. Autom. Cont., Vol. AC-15, pp. 416-426, August, 1970.

[2] Tse, E., "Observer-Estimators for Discrete-Time Systems," IEEE Trans. Autom. Cont., Vol. AC-18, pp. 10-16, February, 1973.

[3] Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, J. Basic Eng., Ser. D, Vol. 82, pp. 35-45, Mar. 1960.

[4] Luenberger, D. G., "Observing the State of A Linear System," IEEE Trans. Mil. Electron., pp. 74-80, Apr. 1964.

[5] Sims, C. S., "Linear Estimation in Discrete Random Parameter Systems," 1973 SWIEEECO Record of Technical Papers, April 4-6, 1973, Houston, Texas.

DISTRIBUTED PARAMETER IDENTIFICATION USING GOURSAT THEORY

ROBERT O. HUGHES
Construction Engineering Research Laboratory
Champaign, Illinois

ABSTRACT

Using the classic Goursat technique for determining the solution of the wave equation with a constant speed of propagation, c, a new method for identifying c is being developed. The algorithm does not require knowledge of initial conditions and uses one fixed and two scanning measurements.

The unforced wave equation with fixed end conditions is expressed as

$$c^2 u_{xx}(x,t) = u_{tt}(x,t) \tag{1}$$

with

$$u(0,t) = u(L,t) = 0 \tag{2}$$

for $0 < x < L$ and $0 \leq t \leq T$, where L is the length of the spatial domain and T is the terminal measurement time.

Consider the following three measurements:

1. $g_1(t)$--a fixed measurement at x_m, $0 < x_m < L$,
2. $g_2(t)$--a scanning measurement starting at x_m and scanning with speed a in a positive x-direction, and
3. $g_3(t)$--same as $g_2(t)$ except in a negative x-direction and with speed b.

The speed of the scanners must be constrained *a priori* by

$$a \leq c, \tag{3}$$

$$b \leq c, \tag{4}$$

and

$$a \neq b. \tag{5}$$

Goursat theory states that the solution of Eq. (1) at a point, say $P(x_m,t_1)$, is uniquely determined if the characteristics are known and solution data is given along the two intersecting curves $g_2(t)$ and $g_3(t)$. The identification algorithm proceeds in the following fashion:

1. Guess a value of c and call it $\tilde{c}$. $\tilde{c}$ then determines the slope of a trial characteristic line.
2. Start at a point P_1 on $g_1(t)$ and perform the Goursat calculation to determine the solution at P_1. Call this solution $\tilde{u}_1$.
3. Let P_2 be $P_1 - \Delta$, where Δ is a small quantity. Calculate $\tilde{u}_2$.
4. Continue process for n points. Form vectors $\underline{\tilde{u}}$ and $\underline{u}$.
5. Adjust $\tilde{c}$ until there is a favorable agreement between $\underline{\tilde{u}}$, the calculated value, and $\underline{u}$, the measured value.

APPLICATION OF STOCHASTIC APPROXIMATION TO PROCESS CONTROL

T. PRASAD
Dept. of Civil Engineering
University of Waterloo
Waterloo, Ontario

N.K. SINHA
Dept. of Electrical Engineering
McMaster University
Hamilton, Ontario

SUMMARY

Application of stochastic approximation to the control of processes with uncertainty has been proposed by Saridis and Lobbia[1] in a recent paper. Their considerations are based on the classical approach to adaptive control, i.e., the characteristics of the plant to be controlled must be determined completely so that the parameters of the adjustable controller can be computed in order to optimize the performance for a given figure of merit. As such, they first develop two stochastic approximation algorithms for the identification of the plant parameters. This is followed by the second step, which is to analyze the sensitivity of the performance index of the controlled system for various "suboptimal" controllers which depend on the parameter estimators.

Although this approach is computationally attractive, it is rather slow. As indicated in the numerical example given by Saridis and Lobbia, even for a first-order system it took more than 100 iterations for the parameter identification to converge to within 10% of the parameter values.

An alternative approach is based on the philosophy that a complete knowledge of the plant is not necessary for optimum adjustment of the controller parameters if the cost function is measured continuously and the parameters of the controller are adjusted until the minimum of the cost function is reached. Sinha, Temple and Rey[2] have recently proposed a vector cost function for adaptive control based on this approach, and efficient gradient methods for the minimization of the norm of this vector cost function have been considered by Sinha[3].

The object of this paper is to discuss the application of stochastic approximation algorithms for determining the parameters of the controller which will minimize such a vector cost function in those cases where only noisy measurements of the cost function are available. For the sake of simplicity, the configuration of the controller is fixed, and only its parameters are determined using an iterative technique.

The main difficulty with stochastic approximation has been that the algorithms converge rather slowly. Sinha and Griscik[4] have proposed a new algorithm which converges much faster. This present paper discussed the application of this algorithm to the problem of adaptive control under conditions of uncertainty. A number of examples compare the effectiveness

of this algorithm and others for such an application.

REFERENCES

1. G.N. Saridis and R.N. Lobbia: PARAMETER IDENTIFICATION AND CONTROL OF LINEAR DISCRETE TIME SYSTEMS, IEEE Trans. on Automatic Control, vol. AC-17, February 1972, pp. 52-60.
2. N.K. Sinha, V. Temple and A.J.C. Rey: A VECTOR COST FUNCTION FOR EFFICIENT ADAPTATION, Int. J. of Control, vol. 16, December 1972, pp. 1107-1120.
3. N.K. Sinha: ADAPTIVE CONTROL WITH INCOMPLETE IDENTIFICATION, Proceedings of the Third IFAC Symposium on Sensitivity, Adaptivity and Optimality, (Ischia, Italy), June 1973, pp. 343-348.
4. N.K. Sinha and M.P. Griscik: A STOCHASTIC APPROXIMATION METHOD, IEE Trans. on Systems, Man and Cybernetics, vol. SMC-1, October 1971, pp. 338-344.

UNBIASED ESTIMATES FOR ON-LINE SYSTEM IDENTIFICATION

ABHIJIT SEN AND NARESH K. SINHA
Group on Simulation, Optimization & Control
and Department of Electrical Engineering
McMaster University, Hamilton, Ontario, Canada

ABSTRACT

A modified recursive pseudoinverse algorithm is presented for the determination of unbiased estimates of the parameters of a discrete time system from the input output data in the presence of considerable noise.

SUMMARY

In an earlier paper[1] a recursive pseudoinverse algorithm was presented for obtaining the least square estimate of parameters of discrete time system. Consider a linear discrete time system described by the difference equation

$$x_k = \sum_{j=0}^{m} a_j r_{k-j} - \sum_{j=1}^{n} b_j x_{k-j} \tag{1}$$

where $r_i = r(iT)$ is the input sequence and $x_i = x(iT)$ is the output sequence The observed output as a result of additive noise ξ_k is given by

$$y_k = x_k + \xi_k \tag{2}$$

Combining equations (1) & (2) one may write

$$\underline{y}_k = A_k \underline{\phi} + \underline{e}_k \tag{3}$$

where

$$\underline{y}_k = [y_1 \ y_2 \ . \ . \ . \ y_k]^T \tag{4}$$

$$\underline{e}_k = [e_1 \ e_2 \ . \ . \ . \ e_k]^T, \text{ and } e_k = \xi_k + \sum_{i=1}^{n} b_i \xi_{k-1} \tag{5}$$

$$\underline{\phi} = [a_o \ a_1 \ . \ . \ a_m \ b_1 \ b_2 \ . \ . \ b_n]^T \tag{6}$$

and

$$A_k = \begin{bmatrix} r_1 & r_o & . \ . \ . & r_{1-m} & -y_o & -y_{-1} & . \ . \ . & -y_{1-n} \\ . & & & & & & & \\ . & & & & & & & \\ . & & & & & & & \\ r_k & r_{k-1} & . \ . & r_{k-m} & -y_{k-1} & -y_{k-2} & . \ . \ . & -y_{k-n} \end{bmatrix} \tag{7}$$

The least squares estimates of $\underline{\phi}$ is given by

$$\underline{\hat{\phi}}_k = A_k^+ \underline{y}_k \tag{8}$$

where A_k^+ is the pseudoinverse of A_k defined as

$$A_k^+ = [A_k^T A_k]^{-1} A_k^T \quad \text{for } k > m+n+1 \tag{9}$$

Clarke[2] estimates the residuals by assuming a model

$$\hat{e}_k = -f_1 \hat{e}_{k-1} - f_2 \hat{e}_{k-2} \ . \ . \ . -f_s \hat{e}_{k-s} + w_k \tag{10}$$

where

$$\hat{e}_k = y_k - \sum_{j=0}^{m} \hat{a}_j \, r_{k-j} + \sum_{j=1}^{n} \hat{b}_j \, y_{k-j} \tag{11}$$

and w_k is assumed to be an uncorrelated random sequence.

From (11) one may write

$$\underline{\hat{e}}_k = G_k \, \underline{\psi} + \underline{w}_k \tag{12}$$

where

$$\hat{e}_k = [\hat{e}_1 \; \hat{e}_2 \; . \; . \; . \; \hat{e}_k]^T \tag{13}$$

$$\underline{\psi} = [f_1 \; f_2 \; . \; . \; . \; f_s]^T \tag{14}$$

$$\underline{w}_k = [w_1 \; w_2 \; . \; . \; . \; w_k]^T \tag{15}$$

and

$$G_k \begin{bmatrix} -\hat{e}_0 & -\hat{e}_{-1} & . \; . \; . & -\hat{e}_{1-s} \\ . & & & \\ . & & & \\ -\hat{e}_{k-1} & -\hat{e}_{k-2} & . \; . \; . & -\hat{e}_{k-s} \end{bmatrix} \tag{16}$$

The least squares estimates of $\underline{\psi}$ is given by

$$\underline{\hat{\psi}}_k = G_k^{+} \, \underline{\hat{e}}_k \tag{17}$$

where G_k^{+} is the pseudoinverse of G_k ie

$$G_k^{+} = [G_k^T \, G_k]^{-1} \, G_k^T \quad \text{for } k > s \tag{18}$$

utilising this estimate $\underline{\hat{\psi}}_k$ the sequences r_k and y_k are filtered to obtain

$$r_k^{*} = r_k + \sum_{j=1}^{s} \hat{f}_j \, r_{k-j} \tag{19}$$

$$y_k^{*} = y_k + \sum_{j=1}^{s} \hat{f}_j \, y_{k-j} \tag{20}$$

From these filtered sequences a corrected matrix A_k^{*} is obtained instead of A_k in equation (3) using r_j^{*}, y_j^{*} instead of r_j and y_j respectively. Equation (3) can be used with A_k^{*} to obtain unbiased estimate.

As an example, the method was applied to first order system described by

$$x_{k+1} = a x_k + b \, r_k \tag{21}$$

$$y_k = x_k + \alpha \zeta_k \tag{22}$$

a and b were taken as 0.5 and 0.7 respectively. Several values of α were taken for setting the noise level. The results of the final estimates using 500 samples were given below

α	Generalised Pseudoinverse		Correlation Method	
	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
0.5	0.485	0.709	0.453	0.799
0.7	0.477	0.713	0.480	0.782

REFERENCES

1. Sinha, N.K. and Pille, W.: "On line parameter estimation using matrix pseudoinverse", Proc. I.EE., vol. 118, August 1971, pp.1041-1046

2. Clarke, D.W.: "Generalised least square estimation of the parameters of a dynamic model", 1st IFAC Symp. on Identification", Prague, 1967. Paper 3.17.

3. R.K. Mehra: "On line identification of linear dynamic systems with application to Kalman filtering", IEEE Trans. on Automatic Control, vol. AC-16, February 1971, pp. 12-21.

INVERTIBILITY CONDITIONS FOR A CLASS OF FINITE STATE SYSTEMS EVOLVING ON GROUPS

ALAN S. WILLSKY
Department of Electrical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts

ABSTRACT

A class of systems, introduced by Brockett and Willsky, that evolve homomorphically on finite groups is considered. A general result, extending the linear sequential circuit Hankel matrix result of Massey and Sain, is derived. Also, an analog of the Brockett-Mesarovic result is obtained when we restrict our attention to abelian group systems.

I. INTRODUCTION

The question of invertibility of a dynamical system — i.e. when does the output sequence (function) uniquely determine the input sequence (function) — has received a good deal of attention in the literature [1]-[6] and has applications to problems in coding theory and functional controllability (the dual of invertibility). Massey and Sain [1],[2], Brockett and Mesarovic [3], and Forney [4] have all dealt with invertibility conditions for linear systems. In this highly structured setting one can obtain quite detailed and explicit invertibility conditions. At the other structural extreme, Olson [5] has obtained invertibility results for general finite state machines. Of course the results in this general setting are not nearly as detailed as the linear results.

Our work falls in between these structural extremes. Brockett and Willsky [6],[7] have introduced a class of finite state systems that evolve homomorphically on finite groups. As their results indicate, this class of systems, although much broader than the class of linear sequential circuits (which is included as a special subclass of the class of finite group systems) possess many of the properties of the more structured class of linear systems. In Section II we review some of the important results from [6],[7] on controllability, observability, minimality, and realizability. Section III contains the analog of the Massey-Sain Hankel matrix result for linear sequential circuits. Our proof parallels that in [1], although we must work somewhat harder because of the more general (i.e. nonabelian) setting. In Section IV we derive the analog of the Brockett-Mesarovic [3] linear system result for the class of abelian group systems.

Our present treatment of these questions is necessarily brief. For a thorough treatment, some examples, and discussions of several related problems (such as the construction of inverses), we refer the reader to [8].

II. A CLASS OF FINITE GROUP SYSTEMS

In this section we review some of the basic definitions and results from [6],[7], and [10]. We first note that to be precise we should denote a group $\mathcal{G}$ by a pair $(G,*)$, where G is a set and $*$ is the group operation assigning to every pair $g_1,g_2 \in G$ the element $g_1 * g_2$. We will abuse this notation by identifying $\mathcal{G}$ with G and by denoting $g_1 * g_2$ by g_1g_2.

Definition 1: Let X,U, and Y be finite state, input and output groups, respectively. The dynamical system

$$x(k+1) = b[u(k)]a[x(k)] \qquad y(k) = c[x(k)] \tag{1}$$

where $a:X \to X$, $b:U \to X$, and $c:X \to Y$ are group homomorphisms, is called a finite group homomorphic sequential system (FGHSS), [6].

The next few results, proven in [6] and [10], reflect the highly structured nature of the class of FGHSS's.

Theorem 1: The input, state, and output of the FGHSS (1) are related by

$$x(k) = b[u(k-1)]ab[u(k-2)] \ . \ . \ .a^{k-1}b[u(0)]a^{k}[x(0)] \qquad (2)$$

$$y(k) = T_{o}[u(k-1)]T_{1}[u(k-2)]. \ . \ .T_{k-1}[u(0)]ca^{k}[x(0)] \qquad (3)$$

where $T_i:U \to Y$ is the homomorphism defined by $T_i = ca^{i}b$. ■

Proof: See [6].

The reader is referred to [6] for a result, much like the linear system result, on when a "weighting pattern" $T_0, T_1, T_2 \ldots$ can be realized as a FGHSS. The reader is also referred to the general definitions of and FGHSS results for controllability, distinguishability, and observability given in [6] and [7].

We now note some of the complications and discrepancies with the results of linear theory, caused by the present more general (non-abelian) setting (see [6] and [7] for details). The set R_k of states reachable from the identity e in k steps need not be a subgroup of X, although the set of K_k of states indistinguishable from e over k steps is a normal subgroup. Here (Ra denotes range, ker denotes kernel)

$$R_k = \text{Ra}\{(u_0,\ldots,u_{k-1}) \longmapsto b(u_{k-1})ab(u_{k-2})\ldots a^{k-1}b(u_0)\} \qquad (4)$$

$$K_k = \ker\{x \longmapsto (c(x),\ldots,ca^{k-1}(x))\} \qquad (5)$$

Thus there need not be a controllable FGHSS realization of a FGHSS, although there always is an observable one. See [6],[7] for the proof of the isomorphism of any two minimal (controllable, observable) FGHSS realizations of a given input-output map.

In [6],[7], and [10] we investigated the use of additional assumptions to obtain more detailed results for FGHSS's. If one assumes that the various groups are abelian, essentially all of the results of linear theory go through. A second less restrictive assumption is that $a:X \to X$ is a normal endomorphism (see [6],[7]).

In this case, as proven in [6],[7], R_k is a subgroup of X for all k, and we have the following result (here card $\triangleq$ cardinality):

Theorem 2: Consider a FGHSS with card X = n. Let

$$n = (p_1)^{k_1}(p_2)^{k_2} . \ . \ .(p_s)^{k_s} \qquad (6)$$

where the p_i are distinct primes and the k_i are positive integers. Define

$$k(n) \triangleq \sum_{i=1}^{s} k_s \qquad (7)$$

Then if R_k is a subgroup for all k (e.g. if a is normal or X is abelian), the set R of states reachable from e at some time is given by

$$R \triangleq \bigcup_{k \geq 0} R_k = R_{k(n)} \qquad (8)$$

Thus the system is controllable if and only if $R_{k(n)} = X$.

Proof: See [10]. ■

The lack of controllability – observability duality for FGHSS's is illustrated by the fact that the observability analog of Theorem 2 is true for arbitrary FGHSS's.

Theorem 3: Consider a FGHSS with card X = n. Then for $k \geq k(n)$, $K_k = K_{k(n)}$. Thus the system is observable if and only if $K_{k(n)} = \{e\}$.

Proof: See [10] ■

We note that the basis of the proofs of Theorems 2 and 3 is Lagrange's Theorem [9]: if H is a subgroup of G($H < G$), card H$\backslash$card G (read: "card H divides card G" (evenly)). We will use this in proving the main result of Section IV.

III. A GENERAL INVERTIBILITY CONDITION FOR FGHSS's

In this section we derive a characterization of invertibility for FGHSS's that is analogous to, and in fact generalizes the linear system result of Massey and Sain [1]. We first define the concept of L-delay invertibility for a general discrete time system.

Definition 2: Consider the general discrete time system

$$x(k+1) = \lambda[x(k),u(k)] \qquad y(k) = \delta[x(k)] \tag{9}$$

where $x \in X$, the state set, $u \in U$, the input set, and $y \in Y$, the output set. Define the input and output vectors

$$U_k = \begin{bmatrix} u(0) \\ \cdot \\ \cdot \\ \cdot \\ u(k) \end{bmatrix} \qquad Y_k = \begin{bmatrix} y(1) \\ \cdot \\ \cdot \\ \cdot \\ y(k) \end{bmatrix} \tag{10}$$

Here $U_k \in U^{k+1}$ (the (k+1) - fold Cartesian product of U with itself) and $Y_k \in Y^k$. We say that the system (9) is invertible with delay L or has an inverse of delay L if (given x(0))U_k can be recovered from Y_{k+L+1} $\forall\, k \geq 0$ – i.e. if Y_{k+L+1} uniquely specifies U_k.

To provide some perspective for our results, we refer the reader to the linear system result of Massey and Sain [1], which is also stated in [8]. The main result of this section is a generalization to FGHSS's of the Massey-Sain result for linear sequential circuits (LSC's) – linear systems over finite fields (see [8]).

Recalling the general invertibility definition, our problem here is the following: consider a FGHSS (1) with $x(0) = e$ (as with LSC's, we can "divide out" the effect of a non-identity initial condition – see (3)), card X = n, card U = m < card Y = p (we need this to have any chance of inverting the system); we wish to find conditions on the system such that an L-delay inverse exists. Defining the input-output homomorphisms $T_i = ca^i b$, we see that we have the relation

$$Y_k = M_k U_{k-1} \tag{11}$$

where $M_k\colon\ U^k \to Y^k$ is defined by

$$\begin{aligned} y(1) &= T_0[u(0)] \\ &\ \vdots \\ y(k) &= T_0[u(k-1)]T_1[u(k-2)]\ldots T_{k-1}[u(0)] \end{aligned} \tag{12}$$

Note that unlike M_k in the LSC case (where it is clear that M_k is a linear operator) and much like the input-state map for FGHSS's, M_k for a FGHSS need not be a homomorphism. However, it always is a homomorphism if Y is abelian.

Theorem 4: A FGHSS has an inverse with delay L if and only if U_0 can be determined from Y_{L+1}.

Proof: The proof of this result is precisely the group analog of the LSC result in [1]. See [7] and [8] for details. ■

Define the maps A_k: $U \to Y^k$, D_k: $U^{k-1} \to Y^k$

$$A_k(u) = (T_0(u), T_1(u), \ldots, T_{k-1}(u)) \tag{13}$$

$$D_k(u_1, \ldots, u_{k-1}) = (e, M_{k-1}(u_1, \ldots, u_{k-1})) \tag{14}$$

and endow Y^k with the direct product group structure. Note that A_k is a homomorphism, D_k need not be, and

$$M_k[u(0), \ldots, u(k-1)] = D_k[u(1), \ldots, u(k-1)]A_k[u(0)] \tag{15}$$

Lemma 1: A FGHSS has an inverse with delay L if and only if

(1) A_{L+1} is one-to-one

(2) If $x \in \text{Ra } A_{L+1}$, $y \in \text{Ra } D_{L+1}$, then $yx \in \text{Ra } A_{L+1} \Rightarrow y = (e,..,e)$, and $yx \in \text{Ra } D_{L+1} \Rightarrow x = (e,..,e)$.

Proof (Necessity): Suppose the FGHSS has an L-delay inverse. Clearly A_{L+1} must be one-to-one, or else there exists an input $u(0) \neq e$ such that $A_{L+1}(u(0)) = (e, \ldots, e)$. Then, from (15)

$$M_{L+1}(u(0), e, \ldots, e) = (e, \ldots, e) = Y_{L+1} = M_{L+1}(e, \ldots, e) \tag{16}$$

so u(0) cannot be recovered from Y_{L+1}. Now suppose we have $x \in \text{Ra } A_{L+1}$, $y \in \text{Ra } D_{L+1}$ such that $yx \in \text{Ra } A_{L+1}$. This implies that there exists an input sequence u(0),...,u(L) and another input u'(0), such that

$$yx = D_{L+1}(u(1), \ldots, u(L))A_{L+1}(u(0)) = A_{L+1}(u'(0)) \tag{17}$$

Since we are assuming L-delay invertibility, we must have u(0) = u'(0), but then we must have y = (e,...,e). Now assume $x \in \text{Ra } A_{L+1}$, $y \in \text{Ra } D_{L+1}$ such that $yx \in \text{Ra } D_{L+1}$. Then there is an input sequence u(0),...,u(L) such that the response to this is the same as the response to e,u'(1),...,u'(L). Thus, by L-invertibility, u(0) = e, and $x = A_{L+1}(u(0)) = (e, \ldots, e)$.

Proof (Sufficiency): We now assume conditions (1) and (2) of the lemma statement hold. Suppose we have two different input sequences u(0),...,u(L) and u'(0),...,u'(L) such that

$$Y_{L+1} = M_{L+1}(u(0), \ldots, u(L)) = M_{L+1}(u'(0), \ldots, u'(L)) = Y'_{L+1} \tag{18}$$

Referring to the definition of M_k in (15), we see that (18) is equivalent to

$$D_{L+1}[u(1), \ldots, u(L)]A_{L+1}[u(0)u'(0)^{-1}] = D_{L+1}[u'(1), \ldots, u'(L)] \tag{19}$$

By condition (2), we must have $A_{L+1}[u(0)u'(0)^{-1}] = (e, \ldots, e)$, and by

(1), $u(0)u'(0)^{-1} = e$ — i.e. $u(0) = u'(0)$. Thus we recover $u(0)$ *uniquely* from Y_{L+1} and our system is L-delay invertible. ■

We refer the reader to the proof of Theorem 4 in [1]. Lemma 1 is the analog of the part of that proof concerning the linear independence of certain columns of the matrix M_L in the LSC case (we were forced to work harder in the present nonabelian setting). We can now prove the analog of the Massey-Sain result.

Theorem 5: A FGHSS has an inverse with delay L if and only if

$$\text{card Ra } M_{L+1} = m \text{ card Ra } M_L \tag{20}$$

where card U = m.

Proof: Note that card Ra D_{L+1} = card Ra M_L. One can show [8] that condition (1) of Lemma 1 is the same as card Ra $A_{L+1} = m$ and condition (2) is equivalent to

$$\text{card } [(\text{Ra } D_{L+1})(\text{Ra } A_{L+1})] = [\text{card Ra } D_{L+1}][\text{card Ra } A_{L+1}] \tag{21}$$

Combining these facts, we see that (20) holds. ■

Theorem 6: Consider a FGHSS with card U = m, card Y = r. Then A_k cannot be one-to-one for

$$k < k_0 = \min (\ell | m\backslash r^{\ell}) \tag{22}$$

and thus $(k_0 - 1)$ is a lower bound on the delay in any inverse of the system.

Proof: Consider $A_k : U \to Y^k$. Then by the First Isomorphism Theorem for groups [9]:

$$U/\ker A_k \simeq A_k(U) < Y^k \tag{23}$$

so

$$m\backslash r^k \text{ card ker } A_k \tag{24}$$

and $\ker A_k = \{e\} \Longleftrightarrow m\backslash r^k$. ■

Note that this result has no analog for LSC's — i.e. there is no lower bound on the delay of an inverse for a LSC. The reason is that in the LSC case $m = q^{s_1}$ and $r = q^{s_2}$, where q is the cardinality of the underlying field and the s_i are the dimensions of the input and output space. Since we must have $m \leq r$, we have $m\backslash r$, so $k_0 = 1$.

IV. AN INVERTIBILITY CONDITION FOR ABELIAN FGHSS's

Sain and Massey [2], Brockett and Mesarovic [3], and Willsky [11] have derived necessary and sufficient conditions of a different form for linear system invertibility; however, we cannot derive an analogous FGHSS result without some restrictions. In this section we derive a result of the desired form for the case of abelian group systems.

We refer the reader to [11] for the proof of a strengthened version of the linear system result derived in [2],[3]. The strengthened result essentially states that a linear system is invertible if and only if it is invertible with delay $L \leq n - m$, where n = dimension of the state space and m = dimension of the input space. Also, a test for invertibility involving a single matrix (as opposed to the full M_k sequence) is derived in [2],[3], and [11].

We now present the abelian FGHSS analogs of these results. Assume that U and Y are abelian groups. In this case, $M_k : U^k$ (direct product) $\rightarrow Y^k$ (direct product) is a homomorphism $\forall$ k, as is $D_k : U^{k-1} \rightarrow Y^k$, defined by (14).

Theorem 7: Consider a FGHSS with U and Y abelian, and consider $A_k : U \rightarrow Y^k$ and $D_k : U^{k-1} \rightarrow Y^k$ defined by (13) and (14) respectively. Then

$$\frac{\text{card Ra } M_{L+1}}{\text{card Ra } M_L} = \frac{\text{card Ra } A_{L+1}}{\text{card}[\text{Ra } A_{L+1} \cap \text{Ra } D_{L+1}]} \tag{25}$$

Proof: We first note that Ra A_{L+1} and Ra D_{L+1} are groups and from (15)

$$\text{Ra } M_{L+1} = \text{Ra } A_{L+1} + \text{Ra } D_{L+1} \tag{26}$$

where we use "+" to denote the commutative group operation on Y^ℓ. Let

$$\text{Ra } A_{L+1} \cap \text{Ra } D_{L+1} = \{z_1, \ldots, z_r\} \tag{27}$$

Choose $w \in \text{Ra } M_{L+1}$ and $w_A \in \text{Ra } A_{L+1}, w_B \in \text{Ra } D_{L+1}$ such that $w = w_A + w_B$. Then

$$w = (w_A + z_i) + (w_B - z_i) \ \forall z_i \in \text{Ra } A_{L+1} \cap \text{Ra } D_{L+1} \tag{28}$$

and if $w = v_A + v_B$, $v_A \in \text{Ra } A_{L+1}$, $v_B \in \text{Ra } D_{L+1}$, then $v_A = w_A + z_i$, $v_B = w_B - z_i$ for any $z_i \in \text{Ra } A_{L+1} \cap \text{Ra } D_{L+1}$. Thus, there are precisely $r = \text{card } [\text{Ra } A_{L+1} \cap \text{Ra } D_{L+1}]$ ways to write each $w \in \text{Ra } M_{L+1}$ as a sum of an element of Ra A_{L+1} and an element of Ra D_{L+1}. Then

$$\text{card Ra } M_{L+1} = \frac{[\text{card Ra } A_{L+1}][\text{card Ra } D_{L+1}]}{\text{card}[\text{Ra } A_{L+1} \cap \text{Ra } D_{L+1}]} \tag{29}$$

and, we observe that card Ra D_{L+1} = card Ra M_L, which proves the theorem. ■

Note that card Ra $A_{L+1} \cap$ Ra D_{L+1} < card Ra A_{L+1} so

$$\frac{\text{card } M_{L+1}}{\text{card } M_L} = \text{an integer} \tag{30}$$

Lemma 2: Consider a FGHSS under the same hypotheses as in Theorem 7. Then the ratio (25) is a nondecreasing function of L.

Proof: The proof of this result is rather long and involves the use of Lagrange's Theorem. We omit the proof here and refer the reader to [7] and [8]. ■

We now wish to bound the maximum possible value for the minimum delay in any inverse of an abelian FGHSS. We put the direct product structure on U^k and Y^k. As proven in [8]

$$\left.\frac{\text{card Ra } M_{k+1}}{\text{card Ra } M_k}\right\backslash m = \text{card } U \tag{31}$$

Suppose no inverse with delay L exists. Then no inverse with delay less than L exists, and therefore

$$\frac{\text{card Ra } M_{k+1}}{\text{card Ra } M_k} < m \quad k = 0,\ldots,L \tag{32}$$

where card $M_0 = 1$. Define

$$s_m = \text{the smallest integer greater than 1 that divides } m \tag{33}$$

$$r_m = \text{the largest integer less than } m \text{ that divides } m (= \frac{m}{s_m}) \tag{34}$$

Combining (31)-(34), we have

$$\text{card Ra } M_{L+1} \leq r_m \text{ card Ra } M_L \leq r_m^2 \text{ card Ra } M_{L-1} \leq \ldots \leq r_m^{L+1} \tag{35}$$

Since $M_{L+1} : U^{L+1} \rightarrow Y^{L+1}$ and card $U^{L+1} = m^{L+1}$, we have

$$\text{card ker } M_{L+1} \geq s_m^{L+1}. \tag{36}$$

We now restrict our attention to ker $M_{L+1} \overset{\Delta}{=} N_{L+1}$. We now assume that $u(0),\ldots,u(L) \in N_{L+1}$ have been applied, and we apply another input $u(L+1) \in U$. The set $N_{L+1} \times U$ of such input strings is a subgroup of U^{L+2} and, referring to (36)

$$\text{card}(N_{L+1} \times U) \geq m s_m^{L+1} \tag{37}$$

Consider the input-state map $E_{L+1} : N_{L+1} \times U \rightarrow X$

$$E_{L+1}(u(0),\ldots,u(L+1)) = b[u(L+1)] + ab[u(L)] + \ldots + a^{L+1}b[u(0)] \tag{38}$$

This is a homomorphism, and, if card$(N_{L+1} \times U) > n =$ card X, E_{L+1} has a nontrivial kernel — i.e. there exists a string $u(0),\ldots,u(L+1)$ not identically zero such that $y(1) = y(2) = \ldots = y(L+1) = e$ $((u(0),\ldots,u(L)) \in N_{L+1})$ and $x(L+2) = e$. Thus the output response to the string $u(0),\ldots,u(L+1),e,e,\ldots,$ is the same as that to the all identity sequence. Thus the system is not invertible. Referring to (37), we see that the hypothesis card$(N_{L+1} \times U) > n$ holds if and only if

$$m s_m^{L+1} \geq n + 1 \tag{39}$$

Thus, the smallest L such that this holds is

$$L_0 = \left\lceil \log_{s_m} \left(\frac{n+1}{m}\right) \right\rceil - 1 \overset{\Delta}{=} q(m, \frac{n+1}{m}) - 1 \tag{40}$$

where $\lceil x \rceil$ = smallest integer $\geq x$. We have proven the following:

Theorem 8: Consider an abelian FGHSS with card U = m, card X = n. Then if the system does not have an inverse with delay L_0, where L_0 is given by (40), it is not invertible. Also the system is L_0 invertible if and only if we have the following condition: given $u(0),\ldots,u(L_0+1)$, such that the output of the system (started at $x(0) = e$) in response to this string followed by all e's is identically e, then $u(0) = \ldots = u(L_0 + 1) = e$. ■

This result is the analog of the strengthened version of the Brockett-Mesarovic-Massey-Sain result [11]. We also have the following results which are the analogs of the corollaries to the linear system result [8], [11].

Corollary 1: An abelian FGHSS is invertible if and only if ker P = {e}, where $P : U^{L_0+1} \rightarrow Y^{L_0+k(n)+1}$ is defined by

$$P[u(0),..,u(L_0+1)] = \begin{bmatrix} cb[u(0)] \\ \vdots \\ ca^{L_0+1}b[u(0)] + ca^{L_0}b[u(1)] + \dots + cb[u(L_0+1)] \\ \vdots \\ ca^{L_0+k(n)}b[u(0)] + \dots + ca^{k(n)-1}b[u(L_0+1)] \end{bmatrix} \tag{41}$$

and k(n) is given by (7). That is, the system is invertible if and only if

$$\text{card Ra } P = mq(m, \frac{n+1}{m}) \tag{42}$$

Proof: By Theorem 8, our system is invertible if and only if the response to $u(0),\dots,u(L_0+1),e,e,\dots$ is not all e's if any of the $u(i) \neq e$. The map P gives the first $L_0 + k(n) + 1$ outputs in response to this sequence. Let x be the state reached from e after the application of $u(0),\dots,u(L_0+1)$. Then the last k(n) outputs in (41) are $c(x), ca(x),\dots,ca^{k(n)-1}(x)$. If these are all equal to e, then, by Theorem 3, $ca^k(x) = e \ \forall k \geq 0$. Thus, we need only check the kernel of P to see if the system is invertible. ■

Corollary 2: An abelian FGHSS is invertible if and only if

$$\text{card Ra } M_{L_0+1} \geq r_m^{L_0+1} + 1 \tag{43}$$

Proof: This follows from (35). ■

We note that one can make comments relating these results to the LSC results much like those comments made in [7] and [10], and we refer the reader to those remarks. Essentially, our results are identical to the LSC results if the cardinality of the underlying field of the LSC is a prime ρ. If it is a prime power ρ^s, the LSC result is stronger, since in the linear case, we *know* that card Ra M_{k+1}/card Ra M_k grows in multiples of the cardinality of the field (i.e. $r_m = \rho$, but we could actually replace it with ρ^s — see [7],[10]).

V. CONCLUSIONS

In this paper we have studied the invertibility properties of a class of systems evolving homomorphically on finite groups. Although this setting is much more general than that for the class of linear sequential circuits, we have been able to derive results that are strikingly similar to those for LSC's. We have obtained the analog of one characterization [1] of LSC invertibility for the full class of FGHSS's, and, after restricting our attention to abelian systems, we have derived the analog of another LSC result [2],[3],[11].

Several problems associated with the lack of a vector space setting for FGHSS's have been encountered, and it is precisely these difficulties that make our results all the more interesting, as we have had to dig quite deeply into the structure of the systems of interest to obtain our results.

The theory of finite groups is so rich that it is strongly felt that the class of invertible FGHSS's is quite large. Noting that the class of convolutional encoders [4] are a relatively small subclass of the class of FGHSS's, it is hoped that our results will lead to the development of new types of sequential coding systems.

ACKNOWLEDGEMENTS

The author wishes to thank Professor R.W. Brockett for his advice during the preparation of this manuscript. This research was conducted in part while the author was a Fannie and John Hertz Foundation Fellow in the Department of Aeronautics and Astronautics at M.I.T. and in part at the Decision and Control Sciences Group of the M.I.T. Electronic Systems Laboratory with partial support provided by AFOSR under grant 72-2273.

REFERENCES

1. J.L. Massey and M.K. Sain, "Inverses of Linear Sequential Circuits," IEEE Trans. Computers, Vol. C-17, April 1968, pp. 330-337.
2. M.K. Sain and J.L. Massey, "Invertibility of Linear Time - Invariant Dynamical Systems," IEEE Trans. on Aut. Control, Vol. AC-14, No. 2, April 1969, pp. 141-149.
3. R.W. Brockett and M.D. Mesarovic, "The Reproducibility of Multivariable Systems," J. Math. Anal. Appl., Vol. 11, July 1965, pp. 548-563.
4. G.D. Forney, Jr., "Convolutional Codes I: Algebraic Structure," IEEE Trans. Inf. Theory, Vol. IT-16, No. 6, Nov. 1970, pp. 720-738.
5. R.R. Olson, On the Invertibility of Finite State Machines, Dept. of Elec. Eng., Univ. of Notre Dame, Tech Rept. No. EE-703, July 23, 1970.
6. R.W. Brockett and A.S. Willsky, "Finite Group Homomorphic Sequential Systems," IEEE Trans. on Aut. Control, Vol. AC-17, No. 4, August 1972, pp. 483-490.
7. A.S. Willsky, Dynamical Systems Defined on Groups: Structural Properties and Estimation, Ph.D. Thesis, Dept. of Aeronautics and Astronautics, M.I.T., Cambridge, Mass., June 1973.
8. A.S. Willsky, "Invertibility of Finite Group Homomorphic Sequential Systems," submitted for publication.
9. J. Rotman, The Theory of Groups: An Introduction, Allyn and Bacon, Inc., Boston, 1965.
10. R.W. Brockett and A.S. Willsky, "On the Structural Properties of a Class of Systems Evolving on Finite Groups," to be submitted for publication.
11. A.S. Willsky, "On the Invertibility of Linear Systems," to be submitted for publication.

SYNTHESIS OF MULTIVARIABLE REGULATORS

B. FRANCIS, O.A. SEBAKHY and W.M. WONHAM*
Department of Electrical Engineering
University of Toronto
Toronto, Canada M5S 1A4

ABSTRACT

Under mild conditions, the equivalence is established between well-posedness of the "regulator problem with internal stability" (RPIS), and the existence of a synthesis of the regulator which is qualitatively insensitive to a wide class of small parameter variations.

INTRODUCTION

In this paper we investigate the "regulator problem with internal stability" (RPIS) introduced in [1], from the viewpoint of well-posedness and genericity in the sense of [2], and of parametrically insensitive implementation. Subject to mild restrictions it is shown that, if and only if RPIS is well-posed, a controller can be synthesized which preserves output regulation and loop stability in the presence of small parameter variations, of a specified type, in controller and plant. Such a "strong" synthesis is achieved by means of a feedback configuration which in general incorporates an invariant, and suitably redundant, copy of the dynamic model adopted for the exogenous reference and disturbance signals which the system is required to handle. The design technique is a systematic development of the method sketched in [1], §6.

In this conference paper limitations of space necessitate the omission of numerous details, and most proofs of our results. These are provided in the report [3], available from the authors on request.

As in [1] we consider the system

$$\dot{x} = Ax + Bu \ , \ y = Cx \ , \ z = Dx \ , \qquad (1)$$

where y is the measured vector and z the vector to be regulated (e.g. tracking error). The system pair (A,B) describes the plant (controllable subsystem) together with the exogenous reference and disturbance signals (e.g. steps, ramps,...) with respect to which control is needed. Let $\mathcal{X}$ (resp. $\mathcal{U}$) denote the vector space of states (resp. controls), $\langle A|\mathcal{B}\rangle \subset \mathcal{X}$ the controllable subspace of the pair (A,B), $\mathcal{N}_C$(resp. $\mathcal{N}_D$) $\subset \mathcal{X}$ the unobservable subspace of the pair (C,A) (resp. (D,A)), and $\mathcal{X}^+(\cdot) \subset \mathcal{X}$ the unstable subspace of the map in $(\cdot)$**. RPIS amounts to finding state feedback $F : \mathcal{X} \to \mathcal{U}$ such that $\operatorname{Ker} F \supset \mathcal{N}_C$ (observability constraint),

$$\mathcal{X}^+(A+BF) \cap (\langle A|\mathcal{B}\rangle + \mathcal{N}_C) \subset \mathcal{N}_C$$

(internal stability) and $\mathcal{X}^+(A+BF) \subset \operatorname{Ker} D$ (output regulation). A constructive solution of RPIS was obtained in [1].

In the following we shall assume without essential loss of generality that the exogenous signals are completely unstable, namely the

---

* This research was partially supported by The National Research Council of Canada, Grant No. A-7399.

** For detailed definitions see [1] and the references given there.

map $\bar{A}$ induced by A in $\bar{\mathcal{X}} \triangleq \mathcal{X}/\langle A|\mathcal{B}\rangle$ satisfies

$$\sigma(\bar{A}) \subset \mathbb{C}^{+} \triangleq \{\lambda\colon \operatorname{Re}\lambda \geq 0\} \ . \tag{2}$$

It is also natural to assume that the pair (C,A) is detectable, namely

$$\mathcal{X}^{+}(A) \cap \mathcal{N}_{C} = 0 \ , \tag{3}$$

since otherwise, by Theorem 4 of [1], RPIS can be reformulated in the factor space $\mathcal{X}/\mathcal{X}^{+}(A) \cap \mathcal{N}_{C}$. We then have the following result, as an immediate formal simplification of Theorem 1 of [1].

Theorem 1. Subject to assumptions (2) and (3), RPIS is solvable if and only if there exists a subspace $\mathcal{V} \subset \mathcal{X}$ such that

$$\mathcal{V} \subset \operatorname{Ker} D \cap A^{-1}(\mathcal{V}+\mathcal{B}) \tag{4}$$

$$\mathcal{V} \cap \langle A|\mathcal{B}\rangle = 0 \tag{5}$$

$$\langle A|\mathcal{B}\rangle + \mathcal{V} = \mathcal{X} \ . \tag{6}$$

We may clearly assume that $D\colon \mathcal{X} \to \mathcal{Z}$ is epic (otherwise replace $\mathcal{Z}$ by $\operatorname{Im} D$); then (4) and (6) imply

$$D\langle A|\mathcal{B}\rangle = \mathcal{Z} \ . \tag{7}$$

From now on we adopt (7) as an assumption.

WELL-POSEDNESS AND GENERICITY

Write $\mathcal{X}_1 \triangleq \langle A|\mathcal{B}\rangle$, $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$, and fix a compatible basis to obtain matrix representations

$$A = \begin{bmatrix} A_1 & A_3 \\ 0 & A_2 \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$$
$$C = [C_1,\ C_2] \ , \qquad D = [D_1,\ D_2] \ , \tag{8}$$

with dimensions $A_1\colon n_1 \times n_1$, $A_2\colon n_2 \times n_2$, $B_1\colon n_1 \times m$, $C_i\colon p \times n_i$ and $D_i\colon q \times n_i$. Here (A_1, B_1) is controllable and A_2 represents the map $\bar{A}$ defined above, so that

$$\sigma(A_2) \subset \mathbb{C}^{+} \ . \tag{9}$$

Also, in view of (7) we shall assume at the outset

$$\operatorname{Rank} D_1 = q \ . \tag{10}$$

Now fix (A_2, C, D) and regard (A_1, A_3, B_1) as a data point $\underline{p}$ in $\mathbb{R}^N$, $N = n_1^2 + n_1 n_2 + n_1 m$. Following [2], we shall say that RPIS is *generically solvable* if it is solvable at all $\underline{p}$ in the complement of a proper algebraic variety in $\mathbb{R}^N$, and that RPIS is *well-posed at* $\underline{p}$ if it is solvable at all points in some open neighborhood of $\underline{p}$.

Of course many other definitions of well-posedness are possible, the idea being to allow for uncertainty about the precise numerical values of various system parameters. We elect to fix C and D because "typically" the variables which can be measured $(y = Cx)$ or which must be regulated

$(z = Dx)$ are defined by the problem coordinatization in a way which is independent of small parameter variations, for instance when y is a position measurement and z a tracking error. Here we neglect any sensor or comparator imprecision. The matrix A_2 embodies the dynamic structure of disturbance and/or reference signals external to the plant: as such it is normally fixed by *a priori* specification of the class of exogenous signals which the regulator is to be designed to handle: steps, ramps and the like. Similarly the controllable subspace X_1 (i.e. the plant state space) is or can be fixed by coordinatization: since (A_1, B_1) is controllable, so are all pairs (A_1', B_1') in some neighborhood of (A_1, B_1) in $\mathbb{R}^{n_1^2+n_1 m}$. On the other hand, by allowing completely free small variations in (A_1, B_1) we tacitly ignore special features of plant structure often fixed by coordinatization, like definitional relations of the type $\xi_2 = \dot{\xi}_1$. More conservatively still we allow free variations in the map A_3 which binds exogenous signals into the plant. While somewhat unrealistic, such latitude is technically convenient, and it will turn out that our results are not unduly restrictive from a practical point of view.

The main results of this section are the following.

Theorem 2. Subject to (2), (3) and (7) [*equivalently (9), (3) and (10)*], *RPIS is well-posed at* (A_1, A_3, B_1) *iff, over the complex field* $\mathbb{C}$,

$$(A-\lambda)(\operatorname{Ker} D \cap \langle A|\mathcal{B}\rangle) + \mathcal{B} = \langle A|\mathcal{B}\rangle \tag{11}$$

for all $\lambda \in \sigma(\bar{A})$ [*or equivalently*

$$\operatorname{rank} \begin{bmatrix} A_1 - \lambda I & B_1 \\ D_1 & 0 \end{bmatrix} = n_1 + q \tag{12}$$

for all $\lambda \in \sigma(A_2)$] .

Corollary. Under the assumptions of Theorem 2, RPIS is generically solvable iff $m \geq q$.

SYNTHESIS

We shall now link the property of well-posedness with the existence of a 'strong' controller. The latter is to have the desirable property that internal stability and output regulation are preserved, when parameters of the plant and controller undergo small variations. In other words, the controller is 'flexible' enough to permit some uncertainty, at the design stage, about values of system parameters, and also to permit (slow) drift (within limits) of these parameters while the system is in operation. It will be no surprise that a feedback configuration is used: what is emphasized is that the feedback compensator includes a (fixed) model of the external dynamics which in general must be reduplicated in a sense made precise in [3].

To present the main ideas most simply, we shall assume that the measured variables y are precisely the regulated variables z and that the system (1) is completely observable from z ; thus

$$C = D \quad , \quad N_D = 0 \ . \tag{13}$$

For a more general treatment see [3]. As explained in [4], the control law $u = Fx$ provided by formal solution of RPIS can be implemented via a dynamic observer. One actually sets

$$u = Fw \;, \tag{14}$$

where $w \in \mathcal{W}$ is the state of the observer, determined by

$$\dot{w} = (A - KD + BF)w + Kz \;. \tag{15}$$

Comparing (1) and (14)-(15) we get that the state estimation error $e \triangleq x - w \in \mathcal{X} \oplus \mathcal{W}$ is given by

$$\dot{e} = (A - KD)e \;; \tag{16}$$

of course, $K: \mathcal{Z} \to \mathcal{W}$ is selected to make $A - KD$ stable.

Consider the observer map

$$T \triangleq A - KD + BF : \mathcal{W} \to \mathcal{W} \;. \tag{17}$$

With mild abuse of notation, the symbols A, D, B, F here denote copies of the corresponding maps defined before, but now with $\mathcal{X}$ replaced by $\mathcal{W}$; context will make the usage clear. Since F solves RPIS, we know that

$$\mathcal{W}^+(A + BF) \subset \operatorname{Ker} D \;; \tag{18}$$

hence T restricts to $\mathcal{W}^+(A + BF)$, giving

$$\begin{aligned} T|\mathcal{W}^+(A + BF) &= (A + BF)|\mathcal{W}^+(A + BF) \\ &\approx \bar{A} \;. \end{aligned} \tag{19}$$

We can therefore write $\mathcal{W} = \mathcal{W}_1 \oplus \mathcal{W}_2$ with $\mathcal{W}_i \approx \mathcal{X}_i$, and choose a basis for $\mathcal{W}$ such that

$$T \sim \begin{bmatrix} T_1 & 0 \\ T_3 & A_2 \end{bmatrix} , \tag{20}$$

where A_2 is a copy of the matrix A_2 in (8) .

The composite system comprising plant, observer and exogenous signal, with state vector

$$x_S \triangleq x_1 \oplus w \oplus x_2 \in \mathcal{X}_S \triangleq \mathcal{X}_1 \oplus \mathcal{W} \oplus \mathcal{X}_2 \tag{21}$$

has differential equation $\dot{x}_S = A_S x_S$ where, by inspection of (8), (14), (15) and (17) ,

$$A_S \triangleq \begin{bmatrix} A_1 & B_1F & A_3 \\ KD_1 & T & KD_2 \\ 0 & 0 & A_2 \end{bmatrix} . \tag{22}$$

It is natural to introduce the 'loop' matrix

$$A_L \triangleq \begin{bmatrix} A_1 & B_1F \\ KD_1 & T \end{bmatrix} . \tag{23}$$

By (17) and the transformation $(x,w) \mapsto (x,e)$,

$$A_S \approx \begin{bmatrix} A+BF & -BF \\ 0 & A-KD \end{bmatrix} \tag{24}$$

and

$$A_L \approx \begin{bmatrix} A_1+B_1F_1 & -B_1F \\ 0 & A-KD \end{bmatrix} . \tag{25}$$

Thus

$$A_L \text{ is stable ,} \tag{26}$$

as it should be; we have $X_S^+(A_S) \approx X_2$; and the regulation condition is equivalent to

$$X_S^+(A_S) \subset \operatorname{Ker}[D_1 \quad 0 \quad D_2] . \tag{27}$$

We define a synthesis of RPIS (under condition (13)) to be a triple (F,T,K) such that, with A_S and A_L defined by (22) and (23) , (26) and (27) are true. With A_2, D and T_3 fixed, A_S is parametrized by points

$$\underline{p}_S = (A_1,\ A_3,\ B_1,\ F,\ T_1,\ K) \in \mathbb{R}^{N_S} \tag{28}$$

where

$$N_S = n_1^2 + n_1n_2 + n_1m + mn + n_1^2 + nq .$$

We shall say that $\underline{p}_S$ is a regular point if (26) and (27) are true throughout some (open) neighborhood of $\underline{p}_S$ in $\mathbb{R}^{N_S}$. Finally, writing $\underline{p} = (A_1, A_3, B_1)$ as before, we call a synthesis (F,T,K) strong at $\underline{p}$ if $\underline{p}_S$ is regular.

Out objective is to show that, in a suitable sense, RPIS admits a strong synthesis. To construct one we shall need, in general, to modify our given RPIS by a process of 'outward extension'. For this, introduce an auxiliary state space X_a and extended state space $X_e \triangleq X \oplus X_a$. Define extensions D_e, A_e, B_e of D, A, B according to

$$\begin{aligned} &D_e : X_e \to Z : x \oplus x_a \mapsto Dx , \\ &A_e : X_e \to X_e : x \oplus x_a \mapsto (Ax + A_{3a}x_a) \oplus A_{2a}x_a , \\ &B_e : U \to X : u \mapsto Bu . \end{aligned} \tag{29}$$

In terms of our previous choice of basis this amounts to writing

$$A_{3a} = \begin{bmatrix} A_{13} \\ A_{23} \end{bmatrix}, \quad A_e = \begin{bmatrix} A_1 & A_3 & A_{13} \\ 0 & A_2 & A_{23} \\ 0 & 0 & A_{2a} \end{bmatrix} , \quad B_e = \begin{bmatrix} B_1 \\ 0 \\ 0 \end{bmatrix} , \tag{30}$$

and

$$D_e = [D_1 \quad D_2 \quad 0] \ . \tag{31}$$

Here $d(X_a)$ and A_{2a}, A_{3a} can be freely chosen.

Given that (D,A) is observable, we shall say that (D_e, A_e, B_e) is an outward extension of (D, A, B) if (29) [or (30)] holds and (D_e, A_e) is observable. It is obvious that if RPIS is solvable for (D_e, A_e, B_e) then it is solvable for (D, A, B) and that RPIS is imbedded in its extended version in a natural way. Thus, under assumption (13), we say that RPIS is strongly solvable (at p) if there is an outward extension of (D, A, B) for which the corresponding (extended) RPIS has a strong synthesis. Our main result is the following.

Theorem 3. Let $C = D$, D be epic, and $N_D = 0$. Then RPIS is strongly solvable at p if and only if RPIS is well-posed at p.

Several features of the strong synthesis constructed in the proof are worth dwelling on, as they embody structural principles of fundamental importance. First, the compensator input is the regulated output z: in other words a bona fide feedback structure is utilized. Secondly, the fixed part of the compensator (A_2), where we do not and cannot permit parameter variations, will be in general a highly redundant, or 'reduplicated' model of the actual dynamics of the exogenous signals which the designer envisages a priori; the idea is sketched in the Appendix, with details given in [3]. This redundancy is the price we pay for insisting on regulation in the presence of a rich class of variations both internal to the feedback loop and also in the binding map which couples the loop to the exogenous signals. The fact that these two structural features, feedback and redundancy, are actually necessary (as well as sufficient) for strong regulation is easy to illustrate by examples, and will be demonstrated in a subsequent article. Finally we remark that the proof of Theorem 3 shows that our strong synthesis admits any and all perturbations in p_S (see (28)) which preserve the stability of A_L and a mild technical condition (see [3]). Under such perturbations the map $A_1 + B_1F_1$ itself may cease to be stable and the compensator may give up its partial role as observer of the plant state.

CONCLUSION

While the role of feedback in reducing sensitivity is well known, the 'internal model principle' with its redundancy feature is less familiar; see, however [5, p.12], [6, p.338] and more recently [7] for suggestive remarks on this theme. The principle deserves further study.

REFERENCES

[1] W.M. Wonham and J.B. Pearson, "Regulation and internal stabilization in linear multivariable systems", Control Systems Report No. 7212, Dept. of Electrical Engineering, University of Toronto, August, 1972, to appear, SIAM J. Control 12 (1), 1974.

[2] E. Fabian and W.M. Wonham, "Generic solvability of the decoupling problem", Control Systems Report No. 7301, Dept. of Electrical Engineering, Univ. of Toronto, January 1973; to appear, SIAM J. Control, 1974.

[3] B. Francis, O.A. Sebakhy and W.M. Wonham, "Synthesis of multivariable regulators", Control Systems Report No. 7312, Dept. of Electrical Engineering, University of Toronto, October, 1973.

[4] W. M. Wonham, "Tracking and regulation in linear multivariable systems", SIAM J. Control 11 (3) , 1973, pp. 424-437.

[5] O.J.M. Smith, "Feedback Control Systems", McGraw-Hill, New York, 1958.

[6] R. Conant, "The information transfer required in regulatory processes", IEEE. Trans. Systems Science and Cybernetics SSC-5 (4), 1969, pp. 334-338.

[7] E.J. Davison, "The output control of linear time-invariant multivariable systems with unmeasurable arbitrary disturbances", IEEE. Trans. Aut. Control AC-17 (5), 1972, pp. 621-630.

APPENDIX

Proof of Theorem 3 (sketch). For the proof we need a lemma which provides an outward extension with suitable technical properties.

Lemma. Let (D,A,B) *be as in (8),* $\mathcal{N}_D = 0$ *, D epic. There exists an outward extension* (D_e,A_e,B_e) *of* (D,A,B) *such that, if* $\bar{A}_e$ *is the map induced by* A_e *in* $\bar{\mathcal{X}}_e \triangleq \mathcal{X}_e/\langle A_e|\mathcal{B}_e\rangle$*, then*

$$d(\mathrm{Ker}(\bar{A}_e - \lambda)) = q \qquad (32)$$

for each $\lambda \in \sigma(\bar{A})$. *Furthermore, if* $k(\lambda)$ *is the dimension of the largest prime cyclic subspace of* $\bar{\mathcal{X}}$ *corresponding to* $\lambda \in \sigma(\bar{A})$ *, then*

$$\mathrm{Ker}(\bar{A}_e - \lambda)^{k(\lambda)-1} = \mathrm{Im}(\bar{A}_e - \lambda) \ . \qquad (33)$$

In matrix terms, the extension is accomplished as follows. For each $\lambda \in \sigma(A_2)$ each Jordan block of A_2 corresponding to λ is enlarged to a Jordan block of size $k(\lambda) \times k(\lambda)$. Then $q - \nu$ additional Jordan blocks each of size $k(\lambda) \times k(\lambda)$ are adjoined, where $\nu \triangleq d(\mathrm{Ker}(A_2-\lambda))$. Finally A_{13} is chosen to preserve observability. It is easily seen that well-posedness is retained after this outward extension.

We will sketch a proof of Theorem 3 assuming for technical simplicity that $\sigma(A_2)$ is real. Complex eigenvalues can be handled similarly (see [3]).

Proof of Theorem 3. The 'only if' statement being trivial, we assume that RPIS is well-posed. According to the Lemma we can and do assume without loss of generality that, for $\lambda \in \sigma(\bar{A})$,

$$d(\mathrm{Ker}(\bar{A} - \lambda)) = d(Z) \ (=q) \ , \qquad (34)$$

and

$$\mathrm{Ker}(\bar{A} - \lambda)^{k(\lambda)-1} = \mathrm{Im}(\bar{A} - \lambda) \ , \qquad (35)$$

where $k(\lambda)$ is the dimension of the largest prime cyclic subspace of $\bar{A}$ corresponding to λ . Now consider a synthesis with composite system map A_S as in (22). By (34) and (35)

$$d(\mathrm{Ker}(A_2 - \lambda)) = q \ , \quad \lambda \in \sigma(A_2) \ , \qquad (36)$$

and

$$\mathrm{Ker}(A_2 - \lambda)^{k(\lambda)-1} = \mathrm{Im}(A_2 - \lambda) \ . \qquad (37)$$

It will be shown that this synthesis, maybe after slight modification, is strong. To this end consider first (20). If $T_3 \neq 0$, perturb T_1 a little to arrange

$$\sigma(T_1) \cap \sigma(A_2) = \phi , \tag{38}$$

and then redefine $\mathcal{W}_1$ to achieve

$$T = \mathrm{diag}[T_1, A_2] \tag{39}$$

on $\mathcal{W} = \mathcal{W}_1 \oplus \mathcal{W}_2$.

Next, with $F = [F_1, F_2]$, we claim that $F_2 : \mathcal{W}_2 \to \mathcal{U}$ can, if necessary, be perturbed a little to achieve

$$(A_1 - \lambda)\mathrm{Ker}D_1 + B_1 F_2 \mathrm{Ker}(A_2 - \lambda) = \mathcal{X}_1 , \quad \lambda \in \sigma(A_2) . \tag{40}$$

To see this, recall from (11) that well-posedness implies

$$(A_1 - \lambda)\mathrm{Ker}D_1 + \mathcal{B}_1 = \mathcal{X}_1 , \quad \lambda \in \sigma(A_2) . \tag{41}$$

Also, as (D,A) is observable so is (D_1, A_1), and therefore

$$(A_1 - \lambda)\mathrm{Ker}D_1 \approx \mathrm{Ker}D_1 , \quad \lambda \in \mathbb{C} . \tag{42}$$

By (10),(40) and (42),

$$d \left[\frac{(A_1 - \lambda)\mathrm{Ker}D_1 + \mathcal{B}_1}{(A_1 - \lambda)\mathrm{Ker}D_1} \right] = n_1 - (n_1 - q) = q . \tag{43}$$

In view of (36) and (43), (40) is true for all $F_2 \in \mathbb{R}^{m \times n_2}$ with the exception of those F_2 which, as points in $\mathbb{R}^{mn_2}$, belong to some proper algebraic variety in $\mathbb{R}^{mn_2}$.

From now on we assume that (38)-(40) hold; and it is clear that we still have that

$$A_L \text{ is stable.} \tag{44}$$

As in (28) let $\underset{\sim}{p}_S$ denote the parameter of the synthesis as now defined. Since (38), (40) and (44) certainly hold throughout a neighborhood of $\underset{\sim}{p}_S$, it is enough to show that these conditions guarantee the regulation property (27). A_L being stable, $\mathcal{X}_S^+(A_S)$ is uniquely operator-isomorphic to $\mathcal{X}_2$:

$$\begin{tikzcd}
\mathcal{X}_S^+(A_S) \arrow[r, "A_S"] \arrow[d, "\approx"] & \mathcal{X}^+(A_S) \arrow[d, "\approx"] \\
\mathcal{X}_2 \arrow[r, "A_2"] & \mathcal{X}_2
\end{tikzcd} \tag{45}$$

So, pick a cyclic prime subspace of $\mathcal{X}_2$ with eigenvalue λ, dimension k, generator g_2 and basis

$$x_{2j} = (A_2 - \lambda)^{j-1} g_2 , \quad j \in \underline{k} . \tag{46}$$

D_1 being epic, fix $x'_{1j} \in \mathcal{X}_1$ such that

$$D_1 x'_{1j} + D_2 x_{2j} = 0 , \quad j \in \underline{k} . \tag{47}$$

Suppose $x''_{1k}, \ldots, x''_{1,j+1} \in \mathrm{Ker}D_1$ and $w_k, \ldots, w_{j+1} \in \mathcal{W}$ are already defined, with

$$\begin{aligned}
w_\ell &\overset{\Delta}{=} w_{1\ell} \oplus w_{2\ell} \\
w_{1\ell} &= 0 \\
w_{2,k+1} &= 0 \\
(A_2 - \lambda) w_{2\ell} &= w_{2,\ell+1}
\end{aligned} \tag{48}$$

for $\ell = k,\dots,j+1$. Write also

$$x_{1\ell} = x'_{1\ell} + x''_{1\ell}, \quad x_{1,k+1} = 0, \tag{49}$$

and

$$w_{2\ell} = w'_{2\ell} + w''_{2\ell} \tag{50}$$

with $w''_{2\ell} \in \operatorname{Ker}(A_2 - \lambda)$. Now consider the equations

$$\begin{aligned}(A_1 - \lambda)x''_{1j} + B_1F_2w''_{2j} &= -(A_1 - \lambda)x'_{1j} - B_1F_2w'_{2j}, \\ &\quad -B_1F_1w_{1j} - A_3x_{2j} + x_{1,j+1}, \end{aligned} \tag{51a}$$

$$(T_1-\lambda)w_{1j} = w_{1,j+1} \tag{51b}$$

$$(A_2-\lambda)w'_{2j} = w_{2,j+1} . \tag{51c}$$

Since by (48)

$$w_{2,j+1} \in \operatorname{Ker}(A_2 - \lambda)^{k-j} \subset \operatorname{Ker}(A_2 - \lambda)^{k-1}, \tag{52}$$

(37) implies that w'_{2j} exists such that (51c) is true. Next by (38) and (48), (51b) is satisfied if $w_{1j} = 0$. With the right side of (51a) thus determined, (40) now guarantees the existence of $x''_{1j} \in \operatorname{Ker}D_1$ and $w''_{2j} \in \operatorname{Ker}(A_2 - \lambda)$ such that (51a) is true. Thus it follows by backwards induction on j that x_{1j}, w_{1} and w_{2j} exist which satisfy (47)-(51) for all $j \in \underline{k}$; in other words for our fixed set of x_{2j} $(j \in \underline{k})$ there exist vectors $x_{1j} \in \mathcal{X}_1$ and $w_j = w_{1j} \oplus w_{2j} \in \mathcal{W}$ such that

$$x_{1j} \oplus w_j \oplus x_{2j} \in \mathcal{X}_S^+(A_S) \cap \operatorname{Ker}[D_1, 0, D_2] . \tag{53}$$

Applying this argument to each prime cyclic component of a Jordan decomposition of $\mathcal{X}_2$, we conclude that

$$\mathcal{X}_S^+(A_S) \subset \operatorname{Ker}[D_1, 0, D_2] \tag{54}$$

as required. The proof is finished.

ON STEADY-STATE CHARACTERISTICS OF LINEAR MULTIVARIABLE SYSTEMS

SHIGEYUKI HOSOE AND MASAMI ITO
Automatic Control Laboratry, Faculty of Engineering
Nagoya University, Nagoya, Japan

ABSTRACT

Type $[\ell_1,\ldots,\ell m]$ multivariable feedback system is defined, and simple method for finding the type of the system is given. This criterion is also used in defining a type of q-input, m-output linear transfer element. Relations between the types of the feedback system and of the transfer element are clarified by examining the controllability characteristics of the error systems.

1. INTRODUCTION

In classical control theory, the steady-state characteristics of single-input, single-output unity feedback systems have been represented by classifying them into " type ℓ " feedback systems. The utility of the type ℓ feedback system concept is that the designer can simply observe the powers of s in the denominator of the open-loop transfer function and know the exact asymptotic behaviors of the closed-loop system to certain polynomial reference inputs, such as the step function or ramp function. As the result of this, this concept also anables us to have simple design techniques of feedback systems with desired steady-state characteristics to the polynomial reference inputs. The purpose of this paper is to extend these results in classical control theory from single-input, single-output systems to multivariable systems.

The above extension was first done by D. M. Wiberg [1], and recently by N. Sandell and M. Athans [2]. They gave the following definition of type L feedback system:

Definition ([1],[2]) The m-input, m-output system of Fig.1 is called a type L system (L=1,2,...), when $\lim_{t\to\infty} e(t)=0$ (resp.:=0) for the reference inputs $r_i=t^{L-1}U(t)e_i$ (resp.:$=t^L U(t)e_i$) for all i=1,2,...,m (resp.: some i), where U(t) is the unit step function and e_i is the unit i-th vector.

In many cases, however, this definition is insufficient in describing the steady-state characteristics of the multivariable systems, because in these systems the individual outputs (representing in general different physical quantities, such as velocity, position or pressure) may exhibit different asymptotic behaviors to the reference inputs, and therefore these characteristics of the individual outputs cannot be represented by a single integer L. Another shortcoming of their results was that fairly tedious numerical operations were needed to find the types of the given feedback systems. In fact, trial and error were required in both [1] and [2], inversions of the matrices containing the variable s were necessary in [1], and the characteristic polynomials of the transfer function matrices had to be found in [2].

In this paper a definition of a type $[\ell_1,\ldots,\ell m]$ feedback system is proposed to represent the steady-state characteristics of the m-input, m-output feedback system of Fig.1, and method for finding the types of the feedback systems is presented. Only numerical operations required in the present method are the computations of the residues of the open-loop transfer function matrix at s=0, and the computations of the ranks of the constant matices. Further in the paper a definition of a type $[\ell_1,\ldots,\ell m]$ transfer element is given to classify the q-input, m-output linear dynamical systems. The characteristics of the type $[\ell_1,\ldots,\ell m]$ transfer element are made clear by relating them with the controllability characteristics of the error system which will be defined here in a more restricted sence than was by S. P. Bhattacharyya and J. B. Pearson [3]. As a result of this it is shown that a type $[\ell_1,\ldots,\ell_m]$ feedback system can be obtained simply by constructing a unity feedback system comprised of the dynamic compensator in cascade with the type $[\ell_1,\ldots,\ell_m]$ transfer element.

2. TYPE $[\ell_1,\ldots,\ell_m]$ FEEDBACK CONTROL SYSTEM

The system considered is indicated in Fig.1, and is assumed to be asymptotically stable when $r(t)\equiv 0$. $G(s)$ represents an open-loop transfer function matrix with dimension $m\times m$ which is assumed to be proper. The output y, the reference input r and the error e are m-vectors.

The next is the extended concept of the type ℓ systems in classical control theory.

Definition 1   A system indicated in Fig.1 is called a type $[\ell_1,\ldots,\ell_m]$ feedback system, if ℓ_i $(i=1,\ldots,m)$ are such nonnegative intergers that for the reference inputs of the form

$$r_i(t) = t^{k_i-1}U(t)e_i \quad ; \; i=1,\ldots,m \tag{1}$$

(i) $\lim_{t\to\infty} e(t) = 0$, when $0\le k_i\le \ell_i$ for all i, and

(ii) $\lim_{t\to\infty} e(t) \neq 0$, when $k_i\ge \ell_i+1$ for some i.

In the above if the condition (ii) is eliminated, the system is called at least a type $[\ell_1,\ldots,\ell_m]$ feedback control system.

In the definition, $U(t)$ is the unit step function, $e_i(i=1,\ldots,m)$ are i-th unit vector, and $t^{-1}\equiv 0$.

(Remark 1)   By the additivity of the linear system, we have $\lim_{t\to\infty} e(t)=0$ for the reference input ;

$$r(t) = [c_1t^{k_1-1},\ldots,c_mt^{k_m-1}]^T \; , \tag{2}$$

where $0\le k_i\le \ell_i$, and c_i are any real constants. From this, it can be said that the outputs $y_i(i=1,\ldots,m)$, where $y=[y_1,\ldots,y_m]^T$, of the type $[\ell_1,\ldots,\ell_m]$ feedback system are asymptotically decoupled for a class of reference inputs (2).

(Remark 2)   The relation between the present definition and the one in [1] and [2] can be represented by the formula

$$L = \min_i\{\ell_1,\ldots,\ell_m\} \quad . \tag{3}$$

As in the case of single-input, single-output systems, the next task is to find a method with which the type of the feedback system can be determined from its open-loop transfer function matrix. In this section, only the result will be stated, and the proof of it will be given in §5.

Theorem 1   Let $G(s)$ be expanded into its partial fractions

$$G(s) = \sum_{j=1}^{p} G_j \,/\, s^j + \tilde{G}(s) \quad , \tag{4}$$

where G_p is nonzero constant matrix, and $\tilde{G}(s)$ is a rational matrix which has no poles at the origin. Let $G_j{}^i$ be matrix obtained from G_j by deleting the i-th row, i.e.,

$$G_j{}^i = [g_{1j}{}^T,\ldots,g_{i-1,j}{}^T,\, g_{i+1,j}{}^T,\ldots,\, g_{m,j}{}^T]^T \quad , \tag{5}$$

and let N and N_i $(i=1,\ldots,m)$ be constant matrices defined by

$$N = \begin{bmatrix} G_1 & G_2 & \cdots & G_p \\ G_2 & G_3 & \cdot\cdot G_p & 0 \\ \vdots & & & \vdots \\ G_p & 0 & \cdots\cdots & 0 \end{bmatrix} , \quad N_i = \begin{bmatrix} G_1{}^i & G_2{}^i & \cdots\cdots & G_p{}^i \\ G_2{}^i & G_3{}^i & \cdot\cdot\ G_p{}^i & 0 \\ \vdots & & & \vdots \\ G_p{}^i & 0 & \cdots\cdots\cdots & 0 \end{bmatrix} \tag{6}$$

Then the system of Fig.1 is a type $[\ell_1,\ldots,\ell m]$ feedback system, where integers ℓ_i are given by

$$\ell_i = \text{rank } N - \text{rank } N_i \ . \tag{7}$$

3. ERROR SYSTEM

Theorem 1 suggests us to introduce the following definition of a type $[\ell_1,\ldots,\ell m]$ transfer element:

Definition 2 Let $H(s)$, which is assumed to be proper, be a transfer function matrix of a q-input, m-output linear dynamical system P, and let ℓ_i $(i=1,\ldots,m)$ be the integers obtained by (4)-(7) in which $G(s)$ is replaced by $H(s)$. Then the dynamical system P is called a type $[\ell_1,\ldots,\ell m]$ transfer element.

(Remark 3) Using this definition, the type $[\ell_1,\ldots,\ell m]$ feedback system can be alternatively defined as follows: Definition 1' The system of Fig.1 is called a type $[\ell_1,\ldots,\ell m]$ feedback system if the closed-loop system is asymptotically stable, and if the open-loop system is a type $[\ell_1,\ldots,\ell m]$ transfer element.

Note that in Definition 2 the numbers of the inputs and the outputs of the dynamical system P are not assumed to be equal, that is, to any proper linear dynamical system can be assigned some definite set of integers ℓ_i ($i=1,\ldots,m$). This indicates that the type $[\ell_1,\ldots,\ell m]$ transfer element concept must have wider applications than that was used in Definition 1'. Therefore it becomes of interest for us to inquire what characteristics the type $[\ell_1,\ldots,\ell m]$ transfer element possesses and what relations can be established between the types of the feedback systems and the types of the transfer elements.Definition 1' gives a partial answer. Complete answers to the questions will be given in §4 and §5. As a preparation for it, in this section the error system, which was introduced by S. P. Bhattacharyya and J. B. Pearson, is re-defined in a restricted sence, and the controllability conditions of such system are examined.

Now let the dynamical system P, whose transfer function matrix is $H(s)$, be described by

$$P \begin{cases} \dot{x} = Ax + Bu , \\ y = Cx , \end{cases} \tag{8}$$

where the state x is an n_1-vector, the input u is a q-vector, and the output y is an m-vector. To simplify the stability considerations, P will be assumed to be controllable and observable throughout the paper.

Suppose that the output y of P is required to follow the reference input r described by

$$C.G \begin{cases} \dot{z} = Fz \\ r = Hz , \end{cases} \tag{9}$$

where the state z is an n_2-vector, and r is an m-vector. (H,F) is assumed to be observable.

One is then concerned with the error e:

$$e = r - y \ . \tag{10}$$

By annexing (9) to (8), and by (10), we have

$$\dot{\xi} = \hat{A}\xi + \hat{B}u \quad , \quad e = \hat{C}\xi \quad , \tag{11}$$

where

$$\xi = \begin{bmatrix} x \\ z \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} A & 0 \\ 0 & F \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad \hat{C} = [-C \; , \; H] \quad . \tag{12}$$

Using (11), the error system can be obtained as follows: Let the rank of the observability matrix of (11) be δ:

$$\delta = \text{rank } [\hat{C}^T \; , \; \hat{A}^T\hat{C}^T \; , \ldots , \; \hat{A}^{T^{n_1+n_2-1}} \hat{C}^T] \; . \tag{13}$$

Clearly, $n_1 \leq \delta \leq n_1+n_2$. Then using the fact that (C,A) is observable, it is easily seen that there exists a full row rank matrix T of the form

$$T = \begin{bmatrix} I_{n_1} & T_{12} \\ 0 & T_{22} \end{bmatrix} \tag{14}$$

of order $\delta \times (n_1+n_2)$ such that

$$T\hat{A} = \bar{A}T, \; \hat{C} = \bar{C}T, \; \bar{B} = TB \tag{15}$$

$$\bar{A} = \begin{bmatrix} A & \alpha_{12} \\ 0 & \alpha_{22} \end{bmatrix} \quad , \quad \bar{C} = \left[-C, \; \gamma_2 \right] \quad , \quad \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix} \tag{16}$$

where α_{12}, α_{22}, and γ_2 are matrices of order $n_1 \times (\delta - n_1)$, $(\delta - n_1) \times (\delta - n_1)$, and $m \times (\delta - n_1)$ respectively, and $(\bar{C}, \bar{A})$ is observable.

Let $\quad \eta = T\xi$. $\hfill (17)$

Then using (11)-(17), we have the observable system:

$$\dot{\eta} = \bar{A}\eta + \bar{B}u \quad , \qquad e = \bar{C}\eta \quad . \tag{18}$$

This system will be called here the error system to (8) and (9).*

From the special forms of $\bar{A}$ and $\bar{B}$, we have the following theorem:

Theorem 2 (i) The error system (18) is controllable if and only if

$$\delta = n_1 \tag{19}$$

(ii) If $\delta = n_1$, the error system (18) becomes

$$\dot{\eta} = A\eta + Bu \; , \quad e = -C\eta \quad . \tag{20}$$

(iii) If $\delta > n_1$, there exists at least one observable mode of (18) which is uncontrollable, and which originates from the set of eigenvalues of F.

(Proof) (i) and (ii) are clear from (16) and (18). To see (iii), notice that the uncontrollable mode originates from the set of eigenvalues of α_{22}. On the other hand, by (15) we have $T_{22}F = \alpha_{22}T_{22}$. This shows α_{22} has at least one common eigenvalue with F, since T_{22} is a nonzero matrix.(6) This proves (iii)

In the next section the controllability condition (19) will be used in a different form, which is stated in the following lemma:

Lemma 1 Let $(\check{C}, \check{A}, \check{B})$ be any controllable but not necessarily observable realization of $H(s) = C(sI-A)^{-1}B$.
Then the error system is controllable, i.e., $\delta = n_1$ if and only if there exists an $n \times n_2$ matrix U such that

$$\check{C}(sI-\check{A})^{-1}U = H(sI-F)^{-1} \tag{21}$$

(Proof) Since $(\check{C}, \check{A}, \check{B})$ is the controllable realization of $C(sI-A)^{-1}B$, there exists a nonsingular matrix V satisfying

$$\check{C}V = [C, 0] \quad , \quad V^{-1}\check{A}V = \begin{bmatrix} A & 0 \\ A_{12} & A_{22} \end{bmatrix} ,$$

---

* If (8) is replaced with its state feedback system, present definition cioncides with the one of [3].

where A_{12} and A_{22} are matrices with appropriate dimensions. Then from this we have identities:

$$\check{C}(sI-\check{A})^{-1}U = \check{C}V\cdot V^{-1}(sI-\check{A})^{-1}V\cdot V^{-1}U$$

$$= [C,0]\begin{bmatrix} sI-A & 0 \\ -A_{12} & sI-A_{22} \end{bmatrix}^{-1} V^{-1}U = C(sI-A)^{-1}W_1 \quad ,$$

where $V^{-1}U=[W_1^T , W_2^T]^T$, and W_1 is a constant $n_1\times n_2$ matrix. These identities show that (21) is satisfied for some U if and only if we have

$$C(sI-A)^{-1}W_1 = H(sI-F)^{-1} \tag{22}$$

for some W_1. Moreover by taking a Laplace transform and its inverse it can be seen that (22) is also equivalent to

$$Ce^{At}W_1 = He^{Ft} \tag{23}$$

Suppose now $\delta=n_1$. Then from (15) and (16), we have $T_{12}F=AT_{12}$ and $H=-CT_{12}$. Therefore

$$He^{Ft} = -CT_{12}[\sum_{i=0}^{\infty} F^i t^i/i!] = -C[\sum_{i=0}^{\infty} A^i t^i/i!]T_{12} = Ce^{At}(-T_{12}) \ ,$$

which is (23). So, we have (21). Conversely suppose (21) holds for some U. Then we have (23). By repeated differentiation of (23) and substitution of t=0. We have $CA^iW_1=HF^i$, $i=0,1,2,\dots$. Therefore

$$n_1\le\delta = \text{rank } [C^T , A^TC^T,\dots] = \text{rank} \left[\begin{array}{c:c:c} C^T & A^TC^T & \cdots \\ \hdashline -H^T & -F^TH^T & \cdots \end{array}\right]$$

$$= \text{rank} \begin{bmatrix} I \\ -W_1{}^T \end{bmatrix} [C^T , A^TC^T ,\dots] \leqq \text{rank } [C^T, A^TC^T ,\dots] = n_1$$

4. TYPES OF THE TRANSFER ELEMENT AND THE CONTROLLABILITY OF THE ERROR SYSTEM

In this section the characteristics of the type $[\ell_1,\dots,\ell m]$ transfer element is made clear by obtaining the controllability conditions of the error system when the reference input (9) is restricted to the class of polynomial functions of t.

Since the reference inputs here are polynomial functions of t, the matrices F and H in (9) are taken to be the forms;

$$F = \text{diag}[F_1,\dots,F_m] \ , \quad H = \text{diag}[H_1,\dots,H_m] \quad , \tag{24}$$

where $k_i\times k_i$ matrix F_i and $1\times k_i$ vector H_i are given by

$$F_i = \underbrace{\begin{bmatrix} 0 & 1 & 0\cdots 0 \\ 0 & 0 & 1\cdots 0 \\ & & \vdots \\ 0 & \cdots\cdots & 1 \\ 0 & \cdots\cdots & 0 \end{bmatrix}}_{\leftarrow\ k_i\ \rightarrow} , \quad H_i = \underbrace{[1,0,\dots,0]}_{\leftarrow\ k_i\ \rightarrow} \qquad \text{if} \quad k_i>0 \tag{25}$$

$$F_i = \phi \qquad , \quad H_i = \phi \quad , \qquad\qquad \text{if} \quad k_i = 0 \tag{26}$$

respectively. Here the notation ϕ indicates that, if $k_i=0$, the columns and rows corresponding to the i-th block F_i are absent in F and the columns corresponding to the i-th block H_i is absent in H.

We have the following theorem.

Theorem 3 Let the dynamical system P be described by (8), and let the reference input r be described by (9) and (24)-(26). Assume that P is a type $[\ell_1,\dots,\ell m]$ transfer element. Then the corresponding error system (18)

is (i) controllable, if $0 \le k_i \le \ell_i$ for all $i=1,2,\dots,m$ and
(ii) uncontrollable, if $k_i \ge \ell_i+1$ for some i.

(Proof) By Lemma 1, the error system is controllable if and only if

$$H(sI-F)^{-1} = \check{C}(sI-\check{A})^{-1}V \tag{21}$$

is satisfied for some V, where $(\check{C}, \check{A}, \check{B})$ is any controllable realization of $H(s)=C(sI-A)^{-1}B$. To prove the theorem, some specific controllable realization is needed. To obtain this, let $H(s)$ be expanded as follows (See (4) of Theorem 1):

$$H(s) = \sum_{j=1}^{p} H_j/s_j + \tilde{H}(s) \quad , \tag{27}$$

where H_p is a constant nonzero matrix and $\tilde{H}(s)$ is a rational matrix which has no poles at the origin. Let $(\tilde{C}, \tilde{A}, \tilde{B})$ be any minimal realization of $H(s)$. Then one of the controllable realization of $H(s)$ is given by

$$A = \left[\begin{array}{cccc|c} 0 & I_q & & \vdots & \\ 0 & 0 & I_q & \vdots & \\ & & & \vdots & 0 \\ & & & I_q\ \vdots & \\ 0 & \cdots & \cdots & 0\ \vdots & \\ \hline & 0 & & & \tilde{A} \end{array}\right] , \quad B = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ I_q \\ \hline \tilde{B} \end{array}\right] \tag{28}$$

$$C = [H_p \ \cdots\ H_1 \ \vdots\ \tilde{C}] \ .$$

Substituting now (24)-(26) and (28) into (21), the result is

$$\begin{aligned} &\mathrm{diag}\,[f_1(s),\dots,f_m(s)] \\ &\quad = [H_p/s,\ H_p/s^2 + H_{p-1}/s,\dots,\sum_{i=1}^{p} H_i/s^i]V_1 + C(sI-A)^{-1}V_2 \end{aligned} \tag{29}$$

where

$$f_i(s) = [1/s,\ 1/s^2,\dots,\ 1/s^{k_i}] \ , \quad V = [V_1{}^T \ ,\ V_2{}^T]^T \ . \tag{30}$$

Then equating the both sides of (29), we have

$$\mathrm{diag}\,[f_1(s),\dots,\ f_m(s)] = [H_p/s,\ H_p/s^2 + H_{p-1}/s,\dots,\sum_{i=1}^{p} H_i/s^i]V_1 \ , \tag{31}$$

$$\tilde{C}(sI-\tilde{A})^{-1}V_2 = 0, \tag{32}$$

since $\tilde{C}(sI-\tilde{A})^{-1}V_2$ has no poles at the origin.
Therefore it has been shown that the error system is controllable if and only if there exist two matrices V_1 and V_2 satisfying (31) and (32) respectively. Equality (32) can be trivially satisfied, since we may put $V_2=0$.

Now let the matrix V be partitioned as

$$V_1 = \begin{bmatrix} V_{11} & \cdots & V_{1m} \\ V_{21} & \cdots & V_{2m} \\ \cdots & \cdots & \cdots \\ V_{p1} & \cdots & V_{pm} \end{bmatrix} , \tag{33}$$

where V_{ji} $(j=1,\dots,p\ ,\ i=1,\dots,m)$ are $q \times k_i$ matrices. Then (31) becomes

$$\begin{bmatrix} 0^T \\ \vdots \\ f_i \\ 0^T \\ \vdots \\ 0^T \end{bmatrix} = \sum_{j=1}^{p} \ \sum_{k=p-j+1}^{p} H_k V_{ji}/s^{k-p+1}, \ i=1,\dots,m \ . \tag{34}$$

Therefore, equating the powers of $(1/s)$ in (34), we have

$$\begin{bmatrix} H_1 & H_2 \cdots & H_p \\ H_2 & H_3 \cdots H_p & 0 \\ \cdots & & \vdots \\ H_{k_i} & H_p & \vdots \\ \cdot\cdot & & \vdots \\ H_p & 0 \cdots\cdots & 0 \end{bmatrix} \begin{bmatrix} V_{p,i} \\ V_{p-1,i} \\ \vdots \\ V_{p-p_i+1,i} \\ \vdots \\ V_{1,i} \end{bmatrix} = \begin{bmatrix} \begin{array}{c:c:c} e_i & & \\ \hdashline & e_i & 0 \\ \hdashline & & \ddots \\ \hdashline 0 & & e_i \\ \hdashline \end{array} \end{bmatrix} \quad , \quad i=1,\ldots,m \ , \tag{35}$$

where e_i, $i=1,\ldots,m$, are i-th columns of I_m (:Compare this with (6)). Thus we have arrived at the following conclusion: The error system is controllable if and only if (35) can be solved for all V_{ji} ($j=1,\ldots,p$, $i=1,\ldots,m$), that is to say, the relations

$$\text{rank } N = \text{ran}[N \vdots E_{ki}] \quad , \quad i=1,\ldots,m \tag{36}$$

are satisfied for all i, where N represents the constant matrix given by (6) in which G_i, $i=1,\ldots,p$, are replaced with H_i, and E_{ki} represents the righthand side of (35).

Then the desired result of the theorem follows immediately from the following lemma:

Lemma 2 Let N and N_i be the matrices defined by (6). Then,

$$\text{rank } N - \text{rank } N_i = \ell_i \tag{37}$$

if and only if

$$\text{rank } N \begin{cases} = \text{rank } [N \vdots E_{ki}] & \text{if } 0 \le k_i \le \ell_i \quad , \\ < \text{rank } [N \vdots E_{ki}] & \text{if } \ell_i < k_i \quad . \end{cases} \tag{38}$$

(Proof) Set

$$M_i = \begin{bmatrix} m_{i1} \\ m_{i2} \\ \vdots \\ m_{ip} \end{bmatrix} = \begin{bmatrix} g_{i1} & g_{i2} \cdots\cdots & g_{ip} \\ g_{i2} \cdots\cdots\cdots & g_{ip} & 0 \\ \vdots & & \vdots \\ g_{ip} & 0 \cdots\cdots & 0 \end{bmatrix} \quad , \tag{39}$$

where g_{ij} are i-th rows of G_j. Then from (37),

$$\text{rank} \begin{bmatrix} N_i \\ M_i \end{bmatrix} - \text{rank } N_i = \ell_i \quad .$$

This shows there exist ℓ_i rows of M_i which are linearly independent of the rows of N_i. These rows are $m_{i1},\ldots,m_{i\ell_i}$, since from the special forms of N, N_i and M_i it can readily be verified that, if m_{ij} is dependent on the set of rows of N_i, $m_{i,j+1}$ is also dependent on the same set of row vectors. From this we have (38). The converse can be proved similarly.

5. UNITY FEEDBACK SYSTEM CONSTRUCTED BY STABILIZING THE TYPE $[\ell_1,\ldots,\ell_m]$ TRANSFER ELEMENT BY OUTPUTS

This section investigates the steady-state characteristics of a linear unity feedback system of Fig.2, where:

(a) P denotes the type $[\ell_1,\ldots,\ell_m]$ transfer element, and its state space description is given by

$$\dot{x} = Ax + Bu \ , \quad y = Cx \quad . \tag{8}$$

(b) Q denotes the dynamic compensator described by

$$\dot{\bar{x}} = \mathcal{A}\bar{x} + \mathcal{B}e \ , \quad u = \mathcal{C}\bar{x} + \mathcal{D}e \quad . \tag{40}$$

Here the coefficient matrices of (40) are so chosen that the resulting closed-loop system becomes asymptotically stable ; i.e., all the eigenvalues of $\begin{bmatrix} A-B\mathcal{D}C, & B\mathcal{C} \\ -\mathcal{B}C & , \mathcal{A} \end{bmatrix}$ have negative real parts. It must be noted that the compensator of this kind can always be constructed by the method of F. M. Brasch and J. B. Pearson$^{(4)}$ or by using the Luenberger's observer$^{(5)}$.

Theorem 4 Suppose P and Q satisfy the conditions stated above. Then the system of Fig.2 is at least a type $[\ell_1,\ldots,\ell_m]$ feedback system. Furthermore, if the transfer function matrix $[\mathcal{C}(sI-\mathcal{A})^{-1}+\mathcal{D}]$ of (40) has no poles at the origin, then the system is exactly a type $[\ell_1,\ldots,\ell_m]$ feedback system.

(Proof) Let the reference input r of Fig.2 be described by (9) and (24)-(26), and set $k_i=\ell_i$ $(i=1,\ldots,m)$ in (24) and (25). Then, since the general solution of these equations is given by

$$[c_{11}+c_{12}t+\cdots+c_{1\ell_1}t^{\ell_1-1},\ \ldots,\ c_{m1}+c_{m2}t+\cdots+c_{m\ell_m}t^{\ell_m-1}]^T, \quad (41)$$

the first assertion will be true if we can show $\lim_{t\to\infty} e(t)=0$ for any initial conditions of x and $\bar{x}$.

Consider now the error system constructed from (8) and (9). Then from the conditions on P and r, and from Theorem 3, this error system is controllable. Therefore from Theorem 2, it can be described by

$$\dot{\eta} = A\eta + Bu\ ,\quad e = -C\eta\ . \quad (42)$$

Then from this and (40) and from the condition to Q, we have $\lim_{t\to\infty} e(t)=0$, which is the desired result.

In order to prove the second assertion of the theorem, Theorem 1 is needed. So, at this point, proof of Theorem 1 should be given.

(Proof of Theorem 1) Regard G(s) as the transfer function matrix of P of Theorem 4, and represent its minimal realization by (8). Then by replacing (40) with u=e and from the proof given as above, we have

$$\lim_{t\to\infty} e(t) = 0\ , \quad (43)$$

when the reference input r of Fig.1 is given by (41). Now assume that the order of the reference input r_i of r exceed ℓ_i-1 for some i;i.e., r is given by (9) and (24)-(26), in which $k_i>\ell_i$ for some i. Then by Theorem 3, the error system corresponding to this reference input and (8) has uncontrollable mode. This mode is unstable, for by Theorem 2 we know that the uncontrollable mode of the error system originates from the set of eigenvalues of F. Therefore the error system has unstable uncontrollable mode which is observable. This mode cannot be affected by the output feedback u=e=r-y, and remains observable. Therefore we have

$$\lim_{t\to\infty} e(t) \neq 0\ ,\ \text{for some } x(0) \text{ and } \bar{x}(0)\ . \quad (44)$$

This and (43) prove Theorem 1.

To return the proof of the second assertion of Theorem 4, set

$$L(s) = G_P(s)G_Q(s) = [C(sI-A)^{-1}B]\cdot[\mathcal{C}(sI-\mathcal{A})^{-1}\mathcal{B}+\mathcal{D}]\ ,$$

which is the open-loop transfer function matrix of the system of Fig.2, and let N^P , N_i^P and N^L , N_i^L be the matrices obtained from (4)-(6) where G(s) in (4)-(6) is replaced with $G_P(s)$, or L(s) respectively.

Then from the conditions on P and Q, the following relations can be obtained (;after a tedious but straightforward computations.)

$$N^L = N^P\cdot K\ ,\quad N_i^L = N_i^P\cdot K\ , \quad (45)$$

where

$$K = \begin{bmatrix} G_Q(0) & 0 & & \\ G_Q(0)/1! & G_Q(0) & & \bigcirc \\ \vdots & \vdots & & \\ \vdots & \vdots & & \\ G_Q(0)/(p-1)! & G_Q(0)/(p-2)! & \cdots\cdots & G_Q(0) \end{bmatrix} \quad (46)$$

$$G_Q(0) = \frac{\partial}{\partial^j s} G_Q(s)\Bigg|_{s=0} \qquad (47)$$

Therefore, assuming that the system of Fig.2 is a type $[\ell_1',\ldots,\ell m']$ feedback system, we have

$$\begin{aligned}\ell_i' &= \text{rank } N^L - \text{rank } N_i{}^L \\ &= \text{rank } N^PK - \text{rank } N_i{}^PK \leq \text{rank } N^P - \text{rank } N_i{}^P = \ell_i\end{aligned} \qquad (48)$$

since $N_i{}^P$ is the submatrix of N^P obtained by deleting several rows of N^P. With this results, and by the proof of the first assertion of the theorem, we have $\ell_i'=\ell_i$. This proves the Theorem.

6. AN EXAMPLE

Consider the plant whose transfer function matrix is given by

$$G_P(s) = \begin{bmatrix} 1/s(s+1) & 1/s+1 \\ 1/s & 4/s(s+2) \end{bmatrix} \qquad (49)$$

, which can be verified to be the type [1,1] transfer element by Definition 2 and Theorem 1. The object is to compensate the plant so that we have type [1,1] feedback system.*

To do this, it is sufficient to construct a compensator Q which satisfies the condition stated in Theorem 4. In the following, such a compensator will be constructed by using the Luenberger's observer.

Let the minimal realization of $G_P(s)$ be represented by

$$\dot{x}_P = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -2 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} x_P + \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & 0 \\ 0 & -1 \end{bmatrix} u \quad , \quad y = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 4 & 0 \end{bmatrix} x_P \qquad (50)$$

, and compute the state feedback $u=Kx_P$ such that the set of the closed-loop poles coincides with the some preassigned set $\Lambda=\{\lambda_1,\ldots,\lambda_4\}$., which may be arbitrarily chosen so long as every λ_i has negative real part and any λ_i with $\ell m\,\lambda_i=0$ appears in Λ in a conjugate pair. In this example, Λ is taken to be $\{-1, -1, -2, -2\}$. Then the state feedback becomes

$$u = \begin{bmatrix} -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} x_P \quad . \qquad (51)$$

Next, construct the observer for (50). Taking the observer poles to be $\{-3, -3\}$, we have the observer equations as follows:

$$\begin{aligned}\dot{\hat{\eta}} &= \begin{bmatrix} -3 & 0 \\ 0 & -3 \end{bmatrix} \eta + \begin{bmatrix} -1 & 3 \\ 1 & -4 \end{bmatrix} u + \begin{bmatrix} 6 & 0 \\ 0 & 3 \end{bmatrix} y \\ \hat{x}_P &= \begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 1/4 & 0 \\ -1/2 & 1/4 \end{bmatrix} \hat{\eta} + \begin{bmatrix} 3 & 0 \\ -2 & 0 \\ -3/4 & 1/4 \\ 3/2 & -3/4 \end{bmatrix} y \quad .\end{aligned} \qquad (52)$$

Then by setting $x_P=\hat{x}_P$ in (51) and by (52), we have

$$\dot{\hat{\eta}} = A_C\hat{\eta} + B_C\, y \quad , \quad u = C_C\hat{\eta} + D_C\, y \quad , \qquad (53)$$

where

$$A_C = \begin{bmatrix} -13/4 & 3/4 \\ 4 & -4 \end{bmatrix}, \ B_C = \begin{bmatrix} 33/2 & -9/4 \\ -12 & 6 \end{bmatrix}, \ C_C = \begin{bmatrix} 2 & 0 \\ -1/2 & 1/4 \end{bmatrix}, \ D_C = \begin{bmatrix} -6 & 0 \\ 3/2 & -3/4 \end{bmatrix}$$

---

* Much interesting will be the case in which it is required to construct the type [ℓ_1 , ℓ_2] feedback system, where both ℓ_1 and ℓ_2, or either of which exceed 1 [7]. A systematic design procedure including these cases will be presented elsewhere.

Finally set

$$\mathcal{A} = Ac \ , \quad \mathcal{B} = -Bc \ , \quad \mathcal{C} = Cc \ , \text{ and } \quad \mathcal{D} = -Dc \ .$$

Then $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ represent the coeficient matrices of the required compensator, which is represented by (40). Using this we have the feedback system with the structure of Fig.2. It is obvious that the resulting closed-loop system is asymptotically stable, since the closed-loop poles are {-1, -1, -2 , -2, -3, -3}. So by Theorem 4, the system is of type [1,1].

7. CONCLUSION

Extending the type ℓ system concept in classical control theory, two definitions of type $[\ell_1,\ldots,\ell m]$ feedback system and the type $[\ell_1,\ldots,\ell m]$ transfer element were proposed, and the numerical method for computing the types was obtained. The characteristics of the type $[\ell_1,\ldots,\ell m]$ transfer element were made clear by developing the controllability conditions of the corresponding error systems. Using this, it was shown that the type $[\ell_1,\ldots,\ell m]$ feedback system can be obtained simply by constructing a unity feedback system comprised of the dynamic compensator in cascade with the type $[\ell_1,\ldots,\ell m]$ transfer element.

Finally, we remark that the extentions of the here obtained results to the investigation of the steady state characteristics of feedback systems subjecting to the vector disturbance inputs can be found in [8].

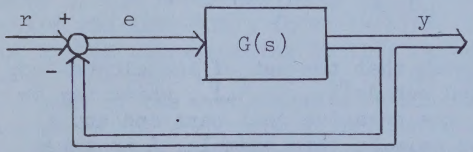


Fig.1 m-input, m-output unity feedback system.

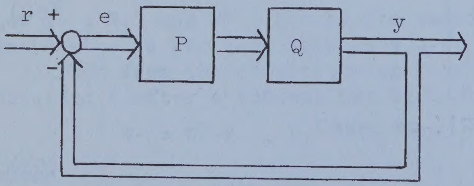


Fig.2 Dynamic compensation of Type $[\ell_1,\ldots,\ell m]$ transfer element.

REFERENCES

1. D.M.Wiberg, State Space and Linear Systems. McGraw-Hill, 1971.
2. N.Sandell and M.Athan,"On Type L Multivariable Linear Systems," Automatica, Vol.9,pp.131-136,1973.
3. S.P.Bhattacharyya and J.B.Pearson, "On Error System and the Servomechanism Problem," INT.J.Control, Vol.15,No.6,pp.1041-1062, 1972.
4. F.M.Brasch and J.B.Pearson,"Pole Placement Using Dynamic Compensator,"IEEE Trans.on A.C., Vol.AC-15 No.1,pp.34-43.
5. D.G.Luenberger,"Observers for Multivariable Systems," IEEE Trans. on A.C.,Vol.AC-11,No.2,pp.190-197, 1966.
6. F.R.Gantmacher, The Theory of Matrices. New York, Chelsea, 1959.
7. S.Hosoe,"Design of the Type $[\ell_1,\ldots,\ell m]$ Feedback Control System," Preprint of the National Convention Record I.E.E.J. in Japan,1973.
8. S.Hosoe, Y.Yonemura and M.Ito, "Steady State Characteristics of Multivariable Feedback Control System With Disturbances," Research Reports of Automatic Control Laboratory, Faculty of Engineering, Nagoya Univ., Vol.20, 1973.

A NEW APPROACH TO SYSTEM REALIZATION FROM INPUT OUTPUT DATA

R.D. GUPTA AND F.W. FAIRMAN
Department of Electrical Engineering
Queen's University
Kingston, Ontario, Canada

ABSTRACT

A method is prsented for estimating the system matrix triple {F,G,H} from noisy measurements of the input and output signals. The algorithm is derived here considering stationary, discrete time systems possessing cyclic state space. Nevertheless, the algorithm is readily applicable to noncyclic systems when viewed together with recent results concerning arbitrary output feedback and cyclicity [4]. The utility of the proposed method rests in avoiding the estimation of a selector matrix as is required in previous methods [1,2].

INTRODUCTION

In the past considerable effort has gone into the development of algorithms for the identification or estimation of single input single output system parameters from noisy measurements of the system signals. The first formulation of an identification algorithm for the multi-input multi-output case was given by Gopinath [1]. An inherent drawback in this approach was the preliminary processing of the input-output data for the construction of a "selector" matrix. Subsequently, a non-iterative version of this procedure was formulated by Budin [2] which made the original algorithm [1] more efficient. However, in practice when the input-output data is corrupted by noise, considerable difficulty is encountered in determining a selector matrix by these procedures. This difficulty has been circumvented in the present method through a formulation which does not require the determination of a selector matrix. This is made possible by restricting the class of systems being considered to those with cyclic state space. These systems have the property that their states can be observed from a linear combination of the output signals.

A controllable and observable system "S" characterized by the matrix triple {F,G,H} and defined by the vector difference equations

$$x(k+1) = Fx(k) + Gu(k) \tag{1}$$

$$y(k) = Hx(k) \tag{2}$$

where F is n x n, G is n x m and H is p x n, is said to be cyclic if its state space is cyclic with respect to F.

The practical justification for the restricted class of systems being considered can be seen by recalling that in order for a system to be noncyclic there must be two identical subsystems having no coupling between them [3]. Occurrences of this type would seem unlikely in practice. Moreover, it is often obvious from physical considerations that the system being identified must be cyclic. However, in cases of structural uncertainty, Davison [4] has suggested a signal processing scheme which extends the present method to include the identification of non-cyclic systems.

Simple manipulation of (1) yields

$$\begin{bmatrix} y(k) \\ y(k+1) \\ \vdots \\ y(k+n-1) \end{bmatrix} = \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix} x(k) + \begin{bmatrix} 0 & 0 \; . \; . \; . & 0 & 0 \\ & 0 \; . \; . \; . & 0 & 0 \\ HG & HG \; . \; . & 0 & 0 \\ HFG & \vdots & \vdots & \vdots \\ \vdots & & & \\ HF^{n-2}G & HF^{n-1}G & HG & 0) \end{bmatrix} \begin{bmatrix} u(k) \\ u(k+1) \\ \vdots \\ u(k+n-1) \end{bmatrix} \tag{3}$$

Next define the n x np matrix

$$S_\alpha = \begin{bmatrix} \alpha' & 0 & \cdots & 0 \\ 0 & \alpha' & & \\ \vdots & & \ddots & \\ 0 & 0 & & \alpha' \end{bmatrix} \tag{4}$$

where α' and 0 and 1 x p constant vectors

$$\alpha' = [\alpha_1 \quad \alpha_2 \; . \; . \; . \alpha_p]$$
$$0 = [0 \quad 0 \; . \; . \; . 0\,]$$

Multiplying (3) throughout by S_α yields

$$\begin{bmatrix} \alpha\ y(k) \\ \alpha\ y(k+1) \\ \vdots \\ \alpha\ y(k+n-1) \end{bmatrix} = \begin{bmatrix} \alpha\ H \\ \alpha\ HF \\ \vdots \\ \alpha HF^{n-1} \end{bmatrix} x(k) + \begin{bmatrix} 0 & 0 & \ldots 0 \\ \alpha\ HG & 0 & \vdots \\ \alpha\ HFG & \alpha\ HG & \\ \vdots & \vdots & \\ \alpha\ HF^{-2} & \alpha\ HF^{n-1}G \ldots & \alpha\ HG \end{bmatrix} \begin{bmatrix} u(k) \\ u(k+1) \\ \vdots \\ u(k+n-1) \end{bmatrix} \tag{5}$$

It is well known, [5], that if the system is cyclic and

$$\text{Rank} \begin{bmatrix} \alpha & H \\ \alpha & HF \\ \alpha & HF^2 \\ \vdots & \\ \alpha & HF^{n-1} \end{bmatrix} = n$$

then there exists a basis for the state space such that

$$\alpha'\ H = [1,\ 0,\ 0\ \ldots,\ 0] \tag{6}$$

and

$$F = \left[\begin{array}{c:c} 0 & I \\ \hline \multicolumn{2}{c}{\alpha'} \end{array}\right] \tag{7}$$

$$a' = [a_n \quad a_{n-1} \ \ldots \ a_1]$$

In view of (6) and (7)

$$\alpha' HF^i G = g'_{i+1}, \qquad i = 0,\ 1,\ \ldots,\ n-1 \tag{8}$$

where g'_i = ith row of G.

From the foregoing (5) can be rewritten as

$$\bar{y}_\alpha(k) = x(k) + R_1\,\bar{u}(k) \tag{9}$$

where

$$R_1 = \begin{bmatrix} 0 & & \cdots 0 & 0 \\ g_1 & & \cdots 0 & 0 \\ g_2 & g_1 & & \\ \vdots & & & \\ g_{n-1} & g_{n-2} & \cdots\ g_1 & 0 \end{bmatrix} \tag{10}$$

and

$$\bar{u}(k) = [u'(k),\ u'(k+1), \ldots,\ u'(k+n-1)]' \tag{11}$$

$$\bar{y}_\alpha(k) = [\alpha' y(k),\ \alpha' y(k+1), \ldots, \alpha' y(k+n-1)]' \tag{12}$$

Substituting x(k) and x(k+1) from (9) into (1) yields

$$\bar{y}_\alpha(k+1) = \begin{bmatrix} F & R \end{bmatrix} \begin{bmatrix} \bar{y}_\alpha(k) \\ \bar{u}(k) \end{bmatrix} \tag{13}$$

where

$$R = R_2 - F\,R_1 \tag{14}$$

$$R_2 = \begin{bmatrix} g_1 & 0 & \cdots & 0 \\ g_2 & g_1 & \cdots & 0 \\ \vdots & & & \vdots \\ g_n & g_{n-1} & \cdots & g_1 \end{bmatrix} \tag{15}$$

Since F is assumed to be in Companion form, it can be easily shown that

$$R = \begin{bmatrix} 0 \\ \hline r' \end{bmatrix} \tag{16}$$

where

$$r' = [r'_n \quad r'_{n-1} \quad \ldots \quad r'_1]$$

$$r'_i = [r_{i1} \quad r_{i2} \quad \ldots \quad r_{im}]$$

and

$$r'_1 = g'_1$$

$$r'_j = g'_j - \sum_{i=1}^{j-1} a_i g'_{j-i} \qquad j = 2,3,\ldots,n \tag{17}$$

Determination of F and R

From the foregoing it is seen that the parameters of F and R appear only in the last scalar equation in (13) i.e.,

$$y_\alpha(k+n) = \phi' \begin{bmatrix} \bar{y}_\alpha(k) \\ \bar{u}(k) \end{bmatrix} \tag{18}$$

where

$$\phi' = [a', r'] \text{ and } y_\alpha(k) = \alpha' y(k)$$

This leads to the identification equation

$$\bar{y}^*_\alpha(k+n) = \phi' D_n(k) \tag{19}$$

where

$$\bar{y}^*_\alpha(k+n) = [y_\alpha(k+n), y_\alpha(k+n+1), \ldots, y_\alpha(k+n^*-1)]$$

and

$$n^* = n + mn$$

$$D_n(k) = \begin{bmatrix} \bar{y}_\alpha(k), & \bar{y}_\alpha(k+1) & \ldots & \bar{y}_\alpha(k+n^*-1) \\ & & & \\ \bar{u}(k), & \bar{u}(k+1) & \ldots & \bar{u}(k+n^*-1) \end{bmatrix} \tag{20}$$

which has a unique solution whenever $D_n(k)$ has an inverse, i.e.

$$\phi = \bar{y}^*(k+n)\, D_n^{-1}(k) \tag{21}$$

Theorem

If S is completely controllable and u(i) are random variables with adjoint non-lattice distributions then $D_n(k)$ is non-singular almost surely.

Proof

Using (9), $D_n(k)$ can be factorized as

$$D_n(k) = \left[\begin{array}{c:c} I_n & R_1 \\ \hdashline 0 & I_{mn} \end{array}\right] \begin{bmatrix} x(k) & x(k+1) & \ldots & x(k+n^*-1) \\ & & & \\ \bar{u}(k) & \bar{u}(k+1) & \ldots & \bar{u}(k+n^*-1) \end{bmatrix}$$

Hence

$$\rho[D_n(k)] = \rho \begin{bmatrix} x(k) & x(k+1) & \ldots & x(k+n^*-1) \\ & & & \\ \bar{u}(k) & \bar{u}(k+1) & \ldots & \bar{u}(k+n^*-1) \end{bmatrix}$$

The remainder of the proof follows from Lemma 2 [1].

Furthermore, it can be shown that for $n_a \geq n$

$$\rho[D_{n_a}(k)] = n + mn_a$$

where n_a is an assumed system order [2]. This result can be used in cases where the system order is not known apriori.

Determination of the Control Matrix G

Having determined ϕ, (17) can be used to generate recursively g_i, the rows of G, i.e.,

$$g_1 = r_1$$

$$g_j = r_j + \sum_{i=1}^{j-1} a_i t_{j-i}, \qquad j = 2,3,\ldots,n \tag{22}$$

Determination of the Observation Matrix H

Using (2) repeatedly enables one to write

$$[y(k)\ y(k+1)\ldots y(k+n^*-1)] = H[x(k), x(k+1), \ldots, x(k+n^*-1)] \tag{23}$$

Substitution of x(k) from (9) in (23) results in

$$y^*(k) = H[I : R_1]\ D_n(k)$$

and H can be determined as

$$H = \begin{array}{l}\text{First } n\\ \text{Columns of}\end{array} \quad \{y^*(k) D_n^{-1}(k)\} \tag{24}$$

where $D_n^{-1}(k)$ was determined earlier for the calculation of ϕ.

ESTIMATOR DEVELOPMENT

From the results just derived it is seen that the matrix triple {F,G,H} can be determined provided the necessary sample values required for the determination of $D_n(k)$ can be accurately obtained and provided $D_n(k)$ has an inverse. A correlation technique described recently in [5] is modified to provide a means for estimating the matrix triple {F,G,H} in the presence of measurement noise. For simplicity of presentation the order of the system is assumed known. Unlike other methods [1,2] the present technique does not require the covariance matrix of the output noise.

Estimation of F and G

Suppose the signals involved in (18) cannot be exactly measured due to the presence of additive noise. Then the measured signals, z(k) and v(k) are related to the signals in (18) by

$$z(k) = y(k) + \mu(k)\ ; \qquad v(k) = u(k) + \gamma(k) \tag{25}$$

where $\mu(k)$ and $\gamma(k)$ are vector valued zero mean ergodic random processes, mutually uncorrelated and each uncorrelated with y(k) and u(k). If the covariance matrix for the input measurement noise is given as

$$E\{\gamma(i)\gamma'(j)\} = B_{|i-j|} \tag{26}$$

then the covariance matrix for the augmented input measurement noise, which results when the measured input is augmented as in the noiseless case (10), can be calculated as

$$E\{\bar{\gamma}(i)\bar{\gamma}'(j)\} = \bar{B}_{|i-j|} \tag{27}$$

where

$$\bar{\gamma}(k) = [\gamma'(k),\ '\ (k+1)\ \ldots\ \gamma'(k+n-1)]'$$

The estimator for ϕ can be developed from (18) as follows.

The use of measured signals, (25) modifies (18) to yield

$$z_\alpha(k+n) = \phi' \begin{bmatrix} \bar{z}_\alpha(k) \\ \bar{v}(k) \end{bmatrix} - \phi' \begin{bmatrix} \bar{\mu}_\alpha(k) \\ \bar{\gamma}(k) \end{bmatrix} + \mu(k+n) \tag{28}$$

Next define a row vector

$$\begin{aligned} p'(k) &= [\bar{v}'_\beta(k-n) \ \vdots \ \bar{v}'\ (k)] \\ &= [\bar{v}'(k-n)S'_\beta \ \vdots \ \bar{v}'\ (k)] \end{aligned} \tag{29}$$

where S is the n x nm matrix

$$S_\beta = \begin{bmatrix} \beta' & 0 & \ldots & 0 \\ 0 & \beta' & \ldots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \ldots & \beta' \end{bmatrix}$$

β' is a 1 x m constant vector and

$$\bar{v}'(k) = [v'(k),\ v'(k+1)\ \ldots\ v'(k+n-1)]$$

Multiplying (28) through by $p'(k)$ and taking the expectation of both sides yields

$$E\{z_\alpha(k+n)p'(k)\} = \phi' E\left\{ \begin{bmatrix} z\ (k) \\ \bar{v}(k) \end{bmatrix} p'(k) \right\} - \phi' E\left\{ \begin{bmatrix} \mu_\alpha(k) \\ \gamma_\alpha(k) \end{bmatrix} p'(k) \right\} + E\{\mu(k)p'(k)\} \tag{30}$$

The assumptions made on the noise processes (25, 27) enable (30) to be written as

$$r_{zv} = \phi'(R_{zv} - P) \tag{31}$$

where

$$r_{zv} = E\{z\ (k+n)p'(k)\}$$

$$R_{zv} = E\left\{ \begin{bmatrix} \bar{z}_\alpha(k) \\ \bar{v}(k) \end{bmatrix} p'(k) \right\}$$

$$P = \left[\begin{array}{c:c} 0_1 & 0_2 \\ \hdashline \bar{B}_n S_\beta & \bar{B}_0 \end{array} \right]$$

0_1 = n x n null matrix

0_2 = n x mn null matrix

Hence an estimate of ϕ' can be determined from (31) whenever $(R_{zv} - P)$ has an inverse. Notice that (31) is composed from the measured input and output signals and from certain elements of the covariance matrix of the input measurement noise only. The covariance matrix for the output measurement noise is not required.

Estimation of H

Recall the output equation

$$y(k) = Hx(k) \tag{2}$$

Substituting for x(k) from (9) yields

$$y(k) = [H \;\vdots\; -HR_1] \;\; m(k) \tag{32}$$

Proceeding as was done to develop the ϕ estimator, H can be estimated from

$$H = \text{first n columns of } \{Q(R_{zv} - P)^{-1}\} \tag{33}$$

where

$$Q = E\{z(k)p'(k)\}$$

Estimator Implementation

From the assumption of ergodicity the ϕ estimator can be rewritten for implementation as

$$\hat{\phi}(N) = r(N)[S(N) - P]^{-1} \tag{34}$$

where

$$r(N) = \frac{1}{N}\sum_{k=1}^{N} z_\alpha(k+n)p'(k)$$

$$S(N) = \frac{1}{N}\sum_{k=1}^{N} \begin{bmatrix} \bar{z}(k) \\ \bar{v}(k) \end{bmatrix} p'(k)$$

Similarly the H estimator may be implemented as

$$\hat{H}(N) = \begin{array}{l}\text{First n} \\ \text{columns of}\end{array} \{Q(N)[S(N) - P]^{-1}\} \tag{35}$$

where

$$Q(N) = \frac{1}{N}\sum_{k=1}^{N} z(k)p'(k)$$

It is easily shown that

$$\underset{N\to\infty}{p\cdot\lim}\ \hat{\phi}(N) = \phi \;;\; \underset{N\to\infty}{p\cdot\lim}\ H(N) = H \tag{36}$$

The estimation of F and G follows from (21,22) and $\hat{\phi}$.

An efficient scheme for on line estimation can be developed from the foregoing as is done in [3].

Consideration of Plant Noise

Suppose the input, v(k), is exactly measurable but there is plant noise, $\gamma(k)$ which can be represented as an additive disturbance such that the plant is excited by u(k) where

$$u(k) = v(k) + \gamma(k) \quad , \quad z(k) = y(k) + \mu(k)$$

As before $\gamma(k)$ and $\mu(k)$, the output measurement noise sequences are mutually uncorrelated and each are uncorrelated with the input v(k) and the output y(k). It can be shown that in this case the estimator equations remain unchanged with the exception that

P· is set to zero, i.e. there is no bias.

CONCLUSIONS

A simple scheme has been presented for system identification. Since a selector matrix is not required, this scheme is especially favourable in the case of noisy data. Moreover, no special difficulty is encountered in the case of dependent outputs. This is important when H is constructed from noisy data.

At present there is no rationale for the choice of α and β. In the absence of measurement noise almost any α and β will suffice. The utility of the wide latitude in the choice of α and β may lie in being able to iteratively improve the system parameter estimates in the presence of measurement noise.

REFERENCES

1 B. Gopinath, "On the identification of linear time-invariant systems from input-output data", Bell Syst. Tech. J., Vol. 48, pp. 1101-1113, May 1969.

2 M.A. Budin, "Minimal realization of discrete linear systems from input-output observations", IEEE Trans. Automat. Contr., Vol. AC-16, pp. 395-401, Oct. 1971.

3 B. Gopinath, "On the control of linear multiple input-output systems" Bell. Syst. Tech. J., Vol. 50, pp. 1063-1081, March 1971.

4 E.J. Davison and S.H. Wang, "Properties of linear time-invariant multi-variable systems subject to arbitrary output and state feedback" IEEE Trans. Automat. Contr., Vol. AC-18, pp. 24-32, Feb. 1973.

5 W.R. Wouters, "On-line identification in an unknown stochastic environment", IEEE Trans. on Sys., Man, and Cyb., Vol. SMC-2, pp. 666-668, Nov. 1972.

6 P.E. Caines, "Minimal realization of transfer function Matrices", Int. J. Control, Vol. 13, pp. 529-547, 1971.

A UNIFIED THEORY OF THE HAMILTONIAN FORMULATION OF SYSTEMS

HSIEN-LU HUANG, MEMBER, IEEE
DEPARTMENT OF ELECTRICAL ENGINEERING
WEST VIRGINIA UNIVERSITY
MORGANTOWN, WEST VIRGINIA 26506

ABSTRACT

> "The thing that has been, it is that which shall be, and that which is done, is that which shall be done, and there is no new thing under the sun."
> . . .Ecclesiastes 1:9

This paper extends the applicability of the classical Euler-Lagrange equations to include dissipative nonreciprocal and nonlinear systems. Hamiltonian dynamics, linear graph theory, and variational principles are correlated, and state equations can be derived from the Hamiltonian equations. In this new formulation the action functional is stationary for the specified classes of systems, and the concept of the traditional "conservative systems" is extended to include dissipative components. The results may be useful in modeling systems, in optimization of systems, etc.

I. INTRODUCTION

It is well-known in classical dynamics and system theory (1-11, 23-25, 28-31, 36) that the Hamiltonian function for nondissipative physical systems without ideal drivers can satisfy the canonical-form Euler-Lagrange equation pair and nondissipative systems are conservative (5, 9, 28) and stationary (10, 14).

In a recent report (27), a new Hamiltonian function was proposed for linear dissipative systems with ideal drivers such that this Hamiltonian satisfies the canonical-form Euler-Lagrange equations, however, it deals only with linear systems. In reference (34) an explicit topological formulation on the Lagrangian and Hamiltonian equations for nonlinear networks was presented, however, its discussion of Hamiltonian formulation is limited to the case of lossless networks. In reference (35) a mixed potential function of content and co-content was proposed to describe a system with gyrators as coupling parts, however, the function is not Hamiltonian.

This paper extends the applicability of the classical Euler-Lagrange equations to include dissipative nonreciprocal and nonlinear systems. The paper deals with Hamiltonian dynamics for physical systems, depends essentially on Linear graph theory (13, 15-22) and variational principles (14, 25, 36), and enables derivation of state equations from the Euler-Lagrange equations. It also resolves an incompatibility between results from calculus of variations applied to dissipative systems with or without ideal drivers. The action functional in this new formulation is stationary for the specified classes of systems, which resolves also a hitherto unsolved dilemma--for linear systems the state equations gave a unique solution, but the action functional was not stationary using previous formulations.

The results give a considerably more satisfying, unified theory of the Hamiltonian formulation of systems, and may be useful in optimization of systems, in modeling systems, etc.

In Section II several definitions and lemmas are given for the Hamiltonian formulation for a general nonreciprocal RC linear and nonlinear network in Section III, with examples considered respectively. Finally, the conclusions are included in Section IV.

For the formulation of the Hamiltonian, H, for a nonreciprocal linear dissipative system, it is shown that if the potential energy function, U, of the system includes energy terms associated with ideal sources, dissipative components, and nonenergic components in addition to the usual energy-storage components, then H, thus derived, satisfies the canonical-form Euler-Lagrange equations. For nonlinear systems the formulation of H is similar to that for the linear systems, except that the relations between voltages and currents are not necessarily linear.

II. DEFINITIONS AND LEMMAS

Def. 2.1 Topological Generalized Coordinates (TGC)

A lumped bilateral physical system with k energy-storage elements associated with k independent initial conditions is said to have k degrees of freedom where k is equal to the number of natural frequencies of the system (17).

Since the number of generalized coordinates is equal to the number of degrees of freedom of the system (8, 9, 24, 38, 39) there are k generalized coordinates, r_1, r_2, ..., r_k associated with the system. However, they are topologically related to one another, and thus are called topological generalized coordinates (TGC).

Among these k TGC's either tree r's or cotree r's are the fundamental TGC's depending on whether r's are across variables or through variables, respectively. In this paper the generalized coordinates r_i's are across variables, thus fundamental TGC's are tree coordinates. The conjugate momenta or generalized momenta p_i's are through variables such that $\dot{p}r$ or $p\dot{r}$ has the dimension of energy.

Def. 2.2 TGC Space, TG Phase Space and TG State Space

(1) TGC Space. Any lumped, bilateral system which is isomorphic (19, 20) to a linear graph will be termed a LB physical system. For a LB system with k degrees of freedom and k TGC's, $r_1(t)$, --, $r_k(t)$, the TGC space is a k-dimensional space in which the r_i are variables.

(2) TG Phase Space. For a dynamic system as in classical mechanics (8) to describe a motion or dynamic behavior completely, in addition to knowing its position and its energy one must also give the velocity components, or the momenta, at each point of the path to fix its direction. The topological generalized (TG) phase space is a 2k-dimensional space for the description of the motion in which TGC i.e., $\underline{r_i}$, and their k corresponding topological generalized momenta (TGM), p_i, are variables (6, 8, 24).

(3) TG State Space. In order to know the state in the TG phase space all 2k variables must be known. However, to define a state of a system in terms of the measurements of practical interest, not all the 2k variables are essential. The topological generalized (TG) state space is a subspace of the TG phase space in which the state vector $[\underline{r}(t)\ \underline{p}(t)]^T$ is defined.

The state of a system is considered to be defined by the state vector, which in turn is defined by the configuration in the TGC space in the sense that when the values of the TGC are known at a given instant, the state of the system is said to be known at that instant (e.g. $\underline{\dot{x}}(t)$ can be found if $\underline{x}(t)$ is known). The dimension of the TG state space is the number of the state variables chosen, and the least number is equal to k [k is equal to the total number of energy storage elements minus the number of independent spring-only cut-sets and the number of independent mass-only tie-sets (or circuits).] The Hamiltonian function, H, which is proved later to be the energy of a system, is a function of TGC and TGM and, hence, a function of the position in the TG state space as well as in the TG phase space.

Def. 2.3 Conservative System (5, 9, 29, 39)

When the forces F acting on a system are such that the work done by them in the passage of the system from a configuration A to the standard reference configuration B is independent of the way in which this passage is carried out, then the system is said to be conservative. If $\nabla \times F = 0$ everywhere in the region of interest, the work done does not depend on the path, but depends only on the starting and end points.

Def. 2.4 Potential Energy U (5, 9, 29, 39)

In the conservative system as defined in Def. 2.3, the work done by the forces in the passage from A to B as mentioned in the same definition is called the potential energy of the system at configuration A with respect to configuration B.

Def. 2.5 Coenergy (5, 25, 34)

The coenergy $\hat{E}$ for a particle in the mechanical system or a capacitor in the electrical system is defined to be

$$\hat{E} = \int \underline{p}^T \, d\dot{\underline{r}} \tag{1}$$

In a linear mechanical system,

$$\hat{E} = \int (\underline{M}\,\dot{\underline{r}})^T d\,\dot{\underline{r}} = \frac{1}{2}\,\dot{\underline{r}}^T \underline{M}\,\dot{\underline{r}} = \frac{1}{2}\,\underline{p}^T\,\underline{D}\,\underline{p} = T \tag{2}$$

Where $\underline{p}$ is the vector mechanical momentum, $\underline{D}$, the inverse of mass $\underline{M}$, $\underline{r}$, the vector coordinate, and T the kinetic energy of the particle. In an electrical system, $\underline{p}$ is chosen to be the electric charge vector $\underline{q}$, and $\underline{D}$, the inverse of the capacitance matrix $\underline{C}$.

Def. 2.6 Lagrangian Function L'

The Lagrangian L' of a system is defined as the difference between coenergy $\hat{E}$ and potential energy U.

Def. 2.7 Sign Convention (27)

For mechanical systems and their analogous electric systems based on the force-current analogy (21) in a tree chosen according to linear graph theory (13, 15, 16, 18), TGM p_i and TGC r_i have the same orientation in tree and opposite directions in cotree. (For mechanical systems without springs, and their analogous electric systems without inductors, TGM p_i and TGC r_i may also have same orientations both in tree and cotree.)

Def. 2.8 Ideal Gyrator (37)

An ideal gyrator is a nonreciprocal two-port network described by the equations

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 & k \\ -k & 0 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$$

The total energy associated with an ideal gyrator is zero, in fact, it is

$$\int (v_1 i_1 + v_2 i_2)dt = \int (k\, i_2 i_1 - k\, i_1 i_2)dt = 0.$$

The ideal gyrator can be used as an impedance inverter. An L-henry inductor can be replaced by a gyrator of gyration constant k and an equivalent capacitor of C' farads, where C' is equal to L/k^2, and C', in farads, k, in ohms, and L, in henries.

Lemma 1

In a system graph, if the original inductance L is in the cotree of a normal tree (15), then upon inversion to an equivalent capacitance C' through an ideal gyrator, the subject cotree edge is replaced by three edges, among which the capacitance C' branch is in the tree, the input edge, GI, and the output edge, GO, of the graph of the two-port gyrator, are in the cotree.

Lemma 2

No matter whether the integral $W = \int_{t_o}^{t} \dot{r}(x)\,\dot{p}(x)\,dx$ is path-independent or path-dependent, where $\dot{r}(t)$ and $\dot{p}(t)$ may be related to each other linearly or nonlinearly, the following relationships hold:

$$\frac{\partial W}{\partial p(t)} = \dot{\gamma}(t)$$

$$\frac{\partial W}{\partial \gamma(t)} = \dot{p}(t) \qquad (4)$$

Proof I: From calculus, if

$$W(y) = \int_{\partial(y)}^{b(y)} f(y, x)dx \qquad (5)$$

then
$$\frac{dW}{dy} = \frac{\partial}{\partial y}\int_{\partial(y)}^{b(y)} f(y, x)dx$$

$$= f(y, x = b(y))\,\frac{\partial b(y)}{\partial y} - f(y, x = \partial(y))\,\frac{\partial a(y)}{\partial y} + \int_{a(y)}^{b(y)} \frac{\partial f(y, x)}{\partial y}dx \qquad (6)$$

Now $W = \int_{t_o}^{t} \dot{r}(x)\,\dot{p}(x)dx$, therefore,

$$\frac{\partial W}{\partial \gamma(t)} = \frac{\partial}{\partial \gamma(t)}\int_{t_o}^{t} \dot{\gamma}(x)\dot{p}(x)dx = \dot{\gamma}(t)\dot{p}(t)\,\frac{\partial t}{\partial \gamma(t)} - \dot{\gamma}(t_o)\dot{p}(t_o)\frac{\partial t_o}{\partial \gamma(t)}$$
$$+ \int_{t_o}^{t} \frac{\partial(\dot{\gamma}(x)\dot{p}(x))}{\partial \gamma(t)}dx \qquad (7)$$

However, t_o is a constant, and $\dot{\gamma}(x)\dot{p}(x)$ is not a function of $\gamma(t)$, therefore,

$$\frac{\partial t_o}{\partial \gamma(t)} = 0 \quad \text{and} \quad \frac{\partial(\dot{\gamma}(x)\dot{p}(x))}{\partial \gamma(t)} = 0,$$

and
$$\frac{\partial W}{\partial \gamma(t)} = \dot{\gamma}(t)\dot{p}(t)\,\frac{1}{\dot{\gamma}(t)} = \dot{p}(t). \qquad (8)$$

Similarly, $\frac{\partial W}{\partial p(t)} = \dot{\gamma}(t)$. (9)

Note the relation between $\dot{\gamma}(t)$ and $\dot{p}(t)$ does not effect the proof given here, therefore, it is true for both linear and nonlinear relationships between $\dot{\gamma}(t)$ and $\dot{p}(t)$.

Proof II: Suppose that the integral $W = \int_{t_o}^{t} \dot{\gamma}(x)\dot{p}(x)dx$ were path-dependent and W were not a point-function of $\gamma(t)$ and $\dot{\gamma}(t)$. There is a way to make $\gamma(t)$ and $\dot{\gamma}(t)$ contribute to the integral W in an explicit way.

1. For a linear trajectory, $\gamma(x)$, according to reference (36), it is found that

$$\gamma(x) = \gamma(t_o) + \frac{\gamma(t) - \gamma(t_o)}{t - t_o}(x - t_o) \qquad (10)$$

as shown in Figure 1, where $\gamma(x)$ is a trajectory from a family of trajectories, passing through the same two end points B and A.

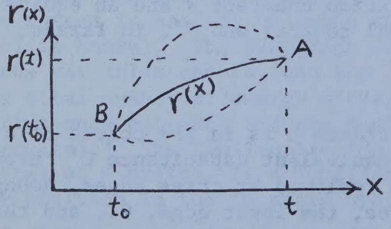


Figure 1

Equation (10) shows that, in this way, $\gamma(x)$ is a point-function of x, $\gamma(t_o)$, $\gamma(t)$, t and t_o.

If $\gamma(x)$ in the integral $\int_{t_o}^{t} \dot{\gamma}(x)\dot{p}(x)dx$ is substituted by Equation (10), and if $\gamma(x)$ is a linear function of p(x), i.e.

$$\gamma(x) = Rp(x) \tag{11}$$

where R is a constant, then

$$W = \int_{t_o}^{t} \dot{\gamma}(x)\dot{p}(x)dx = \int_{t_o}^{t} \left(\frac{\gamma(t) - \gamma(t_o)}{t - t_o}\right)^2 \frac{1}{R}\, dx$$

$$= \frac{1}{R} \frac{(\gamma(t) - \gamma(t_o))^2}{t - t_o} \tag{12}$$

Differentiating W with respect to t, and remembering that $\partial t/\partial\gamma(t) = 1/\dot{\gamma}(t)$, one obtains

$$\frac{\partial W}{\partial \gamma(t)} = \frac{1}{R}\frac{2(\gamma(t) - \gamma(t_o))}{t - t_o} - \frac{1}{R}\frac{(\gamma(t) - \gamma(t_o))^2}{(t - t_o)^2}\; \frac{1}{\dot{\gamma}(t)} = \dot{p}(t) \tag{13}$$

If $\gamma(x)$ is a nonlinear function of $p(x)$, e.g.

$$\gamma(x) = (p(x))^m \triangleq p^m(x) \tag{14}$$

Then

$$p(x) = \gamma^{\frac{1}{m}}(x), \quad \text{and} \quad \dot{p}(x) = \frac{1}{m}\, r^{\frac{1-m}{m}}(x)\, \dot{\gamma}(x).$$

Since $\dot{\gamma}(x) = \dfrac{\gamma(t) - \gamma(t_o)}{t - t_o}$, one obtains

$$W = \int_{t_o}^{t} \dot{\gamma}(x)\dot{p}(x)dx = \int_{t_o}^{t} \dot{\gamma}(x)^2 \frac{1}{m}\gamma^{\frac{1-m}{m}}(x)dx$$

$$= \frac{1}{m}\int_{t_o}^{t} \left(\frac{\gamma(t) - \gamma(t_o)}{t - t_o}\right)^2 \left[\gamma(t_o) + \frac{\gamma(t) - \gamma(t_o)}{t - t_o}(x - t_o)\right]^{\frac{1-m}{m}} dx$$

$$= \frac{\gamma(t) - \gamma(t_o)}{t - t_o}\left\{\gamma^{\frac{1}{m}}(t) - \gamma^{\frac{1}{m}}(t_o)\right\} \tag{15}$$

Differentiating W with respect to $\gamma(t)$ one obtains:

$$\frac{\partial W}{\partial \gamma(t)} = \frac{1}{m}\, \gamma^{\frac{1-m}{m}}(t)\, \dot{\gamma}(t) = \dot{p}(t). \tag{16}$$

and $$\frac{\partial W}{\partial \dot{\gamma}(t)} = 0 \tag{17}$$

Similarly, $$\frac{\partial W}{\partial p(t)} = \dot{\gamma}(t). \tag{18}$$

2. For a parabolic trajectory

$$\gamma(x) = \frac{\gamma(t_o)\, t^2 - \gamma(t)\, t_o^2}{t^2 - t_o^2} + \frac{\gamma(t) - \gamma(t_o)}{t^2 - t_o^2}\, x^2 \tag{19}$$

it can be shown by a similar procedure that

$$\frac{\partial \int_{t_o}^{t} \dot{\gamma}(x)\dot{p}(x)dx}{\partial \gamma(t)} = \dot{p}(t) \tag{20}$$

and $$\frac{\partial W}{\partial p(t)} = \dot{\gamma}(t) \tag{21}$$

Proof III: Let $W = \int_{t_o}^{t} \dot{\gamma}(x)\dot{p}(x)dx$,

$$\frac{\partial W}{\partial \gamma(t)} = \frac{\partial W}{\partial t} \cdot \frac{\partial t}{\partial \gamma(t)} = \frac{\partial W}{\partial x}\frac{\partial x}{\partial \gamma(x)}\bigg|_{x = t}$$

$$= \dot{\gamma}(x)\dot{p}(x)\,\frac{1}{\dot{\gamma}(x)}\Bigg|_{x\,=\,t} \tag{22}$$

$$= \dot{p}(t)$$

Proof IV: $$\frac{\partial W}{\partial\gamma(t)} = \frac{\partial W}{\partial\gamma(x)}\Bigg|_{x\,=\,t} = \frac{\partial\int_{t_o}^{t}\dot{\gamma}(x)\dot{p}(x)dx}{\partial\gamma(x)}\Bigg|_{x\,=\,t}$$

$$= \frac{\partial\int_{t\,=\,t_o}^{t\,=\,t}\dot{p}(x)d\gamma(x)}{\partial\gamma(x)}\Bigg|_{x\,=\,t} = \dot{p}(x)\Bigg|_{x\,=\,t} = \dot{p}(t) \tag{23}$$

Lemma 3

The integral $W = \int_{t_o}^{t}\dot{\gamma}(x)\dot{p}(x)dx$ is a point function of $\gamma(t)$, $\dot{\gamma}(t)$, and t where p may be $M\dot{\gamma}$ or $\gamma(t)$ and $p(t)$ may be related to each other either linearly or nonlinearly.

Lemma 4

Consider a general lumped nonlinear RLC electrical network with gyrators. Let there be n_V voltage-sources, n_I current sources, n_C tree capacitors, n_L cotree inductors, n_G cotree gyrator input edges and n_G gyrator output edges, n_{RT} tree resistors, and n_{RC} cotree resistors. Let the across variables γ_i, or λ_i, associated with independent energy-storage elements be generalized coordinates and the through variables, p_i, or q_i, associated with energy-storage elements be generalized momenta. Both γ and p are identified with appropriate subscripts, e.g., γ_C represents tree capacitance generalized coordinates, p_C, tree capacitance generalized momenta, etc. For a given formulation tree chosen according to theoretic network analysis (15) the circuit and cut-set equations are respectively,

$$\begin{array}{cccccccc} \gamma_V & \gamma_C & \gamma_{RT} & \gamma_L & \gamma_{GI} & \gamma_{GO} & \gamma_{RC} & \gamma_I \end{array}$$

$$\begin{bmatrix} B_{LV} & B_{LC} & B_{LRT} & U & 0 & 0 & 0 & 0 \\ B_{GIV} & B_{GIC} & B_{GIRT} & 0 & U & 0 & 0 & 0 \\ B_{GOV} & B_{GOC} & B_{GORT} & 0 & 0 & U & 0 & 0 \\ B_{RCV} & B_{RCC} & B_{RCRT} & 0 & 0 & 0 & U & 0 \\ B_{IV} & B_{IC} & B_{IRT} & 0 & 0 & 0 & 0 & U \end{bmatrix} \begin{bmatrix} \dot{\gamma}_V \\ \dot{\gamma}_C \\ \dot{\gamma}_{RT} \\ \dot{\gamma}_L \\ \dot{\gamma}_{GI} \\ \dot{\gamma}_{GO} \\ \dot{\gamma}_{RC} \\ \dot{\gamma}_I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{24}$$

and

$$\begin{array}{c}\begin{array}{cccccccc} p_V & p_C & p_{RT} & p_L & p_{GI} & p_{GO} & p_{RC} & p_I \end{array} \\ \begin{bmatrix} U & 0 & 0 & A_{VL} & A_{VGI} & A_{VGO} & A_{VRC} & A_{VI} \\ 0 & U & 0 & A_{CL} & A_{CGI} & A_{CGO} & A_{CRC} & A_{CI} \\ 0 & 0 & U & A_{RTL} & A_{RTGI} & A_{RTGO} & A_{RTRC} & A_{RTI} \end{bmatrix}\end{array} \begin{bmatrix} \dot{p}_V \\ \dot{p}_C \\ \dot{p}_{RT} \\ \dot{p}_L \\ \dot{p}_{GI} \\ \dot{p}_{GO} \\ \dot{p}_{RC} \\ \dot{p}_I \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad (25)$$

where the submatrices B_{IV} is a n_I-by-n_V matrix, A_{VI}, a n_V-by-n_I matrix, etc. (25) Using the meaning of $\frac{\partial \underline{\gamma}_L}{\partial \underline{\gamma}_C}$ given in (25) and taking partial differentiation of γ_I with respect to γ_C, from the last row of the circuit matrix, which states as follows:

$$\underline{\gamma}_I = -B_{IV}\,\underline{\gamma}_V - B_{IC}\,\underline{\gamma}_C - B_{IRT}\,\gamma_{RT} = [\gamma_{I_1}, \gamma_{I_2} \ldots \gamma_{I_{n_I}}]^T, \qquad (26)$$

one obtains

$$\frac{\partial \underline{\gamma}_I}{\partial \underline{\gamma}_C} = -B_{IC} \qquad (27)$$

because co-tree across variables are linear functions of tree across variables. Similarly,

$$\frac{\partial \underline{p}_C}{\partial \underline{p}_I} = -A_{CI}, \text{ etc.} \qquad (28)$$

Furthermore, it can be shown, because of the sign convention chosen

$$A = +B^T. \qquad (29)$$

III. A GENERAL RLC NETWORK

For a dissipative RLC network with ideal sources and nonreciprocal elements, gyrators, the work done in the TGC space by the generalized force $\dot{\underline{p}}_C$ from point A(t) to point B(t) is equal to, using equations (24-29),

$$\begin{aligned}\int_{A(t)}^{B(t)} \dot{\underline{p}}_C^T\, d\underline{\gamma}_C &= \int_t^{t_o} [-A_{CL}\dot{\underline{p}}_L - A_{CGI}\dot{\underline{p}}_{GI} - A_{CGO}\dot{\underline{p}}_{GO} - A_{CRC}\dot{\underline{p}}_{RC} - A_{CI}\dot{\underline{p}}_I]^T \dot{\underline{\gamma}}_c\, dt \\ &= \int_t^{t_o} [\dot{\underline{p}}_{GI}^T(\dot{\underline{\gamma}}_{GI} + B_{GIV}\,\dot{\underline{\gamma}}_V + B_{GIRT}\,\dot{\underline{\gamma}}_{RT}) + \\ &\qquad \dot{\underline{p}}_{GO}^T(\dot{\underline{\gamma}}_{GO} + B_{GOV}\dot{\underline{\gamma}}_V + B_{GORT}\,\dot{\underline{\gamma}}_{RT}) + \\ &\qquad \dot{\underline{p}}_{RC}^T(\dot{\underline{\gamma}}_{RC} + B_{RCV}\,\dot{\underline{\gamma}}_V + B_{RCRT}\,\dot{\underline{\gamma}}_{RT}) + \\ &\qquad \dot{\underline{p}}_I^T\,(\dot{\underline{\gamma}}_I + B_{IV}\,\dot{\underline{\gamma}}_V + B_{IRT}\,\dot{\underline{\gamma}}_{RT}) + \\ &\qquad \dot{\underline{p}}_L\,(\dot{\gamma}_L + B_{LV}\,\dot{\gamma}_V + B_{LRT}\,\dot{\underline{\gamma}}_{RT})]dt \\ &= \int_t^{t_o} (\dot{\underline{p}}_{GI}^T\dot{\underline{\gamma}}_{GI} + \dot{\underline{p}}_{GO}^T\dot{\underline{\gamma}}_{GO} + \dot{\underline{p}}_L^T\dot{\underline{\gamma}}_L + \dot{\underline{p}}_{RC}^T\dot{\underline{\gamma}}_{RC} + \dot{\underline{p}}_I^T\dot{\underline{\gamma}}_I - \dot{\underline{p}}_V^T\dot{\underline{\gamma}}_V \\ &\qquad - \dot{\underline{p}}_{RT}^T\dot{\underline{\gamma}}_{RT})dt = U(t) - U(t_o) \end{aligned} \qquad (30)$$

Since $\frac{\partial \dot{p}_{ci}}{\partial \gamma_{cj}} = 0$ the integral is independent of path in the TGC space, and the system is conservative by the Definition 2.4, the potential energy function is of the form,

$$U = -\int \dot{\underline{p}}_{GI}^T \, d\underline{r}_{GI} - \int \dot{\underline{p}}_{GO}^T \, d\underline{r}_{GO} - \int \dot{\underline{p}}_{L}^T \, d\underline{r}_{L}$$
$$- \int \dot{\underline{p}}_{RC}^T \, d\underline{r}_{RC} - \int \dot{\underline{p}}_{I}^T \, d\underline{r}_{I}$$
$$+ \int \dot{\underline{r}}_{V}^T \, d\underline{p}_{V} + \int \dot{\underline{r}}_{RT}^T \, d\underline{p}_{RT} \qquad (31)$$

Since the coenergy for the system is $\hat{E} = \int \underline{p}_C^T d\dot{\underline{\gamma}}_C$ (32)

the Lagrangian is

$$L' \equiv \hat{E} - U = + \int \underline{p}_C^T d\dot{\underline{\gamma}}_C + \int \dot{\underline{p}}_{GI}^T \, d\underline{r}_{GI} + \int \dot{\underline{p}}_{GO}^T \, d\underline{r}_{GO}$$
$$+ \int \dot{\underline{p}}_{RC}^T \, d\underline{r}_{RC} + \int \dot{\underline{p}}_{I}^T \, d\underline{r}_{I} + \int \dot{\underline{p}}_{L}^T \, d\underline{r}_{L}$$
$$- \int \dot{\underline{r}}_{V}^T \, d\underline{p}_{V} - \int \dot{\underline{r}}_{RT}^T \, d\underline{p}_{RT} \qquad (33)$$

Based on Lemma 3, each term in the right hand side of Eq.(33) is implicitly a point function of $\underline{\gamma}_C$ and $\dot{\underline{\gamma}}_C$, therefore, the lagrangian L' itself is a point function of $\underline{\gamma}_C$, $\dot{\underline{\gamma}}_C$ and t, namely, $L' = L'(\underline{\gamma}_C, \dot{\underline{\gamma}}_C, t)$. (34)

Taking partial differentiation of L' with respect to $\dot{\underline{\gamma}}_C$, one obtains

$$\frac{\partial L'}{\partial \dot{\underline{\gamma}}_C} = \frac{\partial \hat{E}}{\partial \dot{\underline{\gamma}}_C} = \underline{p}_C \qquad (35)$$

Taking partial differentiation of L' with respect to $\underline{\gamma}_C$, the tree coordinate independent variables, and remembering that all cotree coordinates are linear functions of tree coordinates, one obtains, by using Lemmas 2 and 4,

$$\frac{\partial L'}{\partial \underline{\gamma}_C} = - \frac{\partial U}{\partial \underline{\gamma}_C}$$
$$= + \left(\frac{\partial \underline{\gamma}_I}{\partial \underline{\gamma}_C}\right)^T \dot{\underline{p}}_I + \left(\frac{\partial \underline{\gamma}_{RC}}{\partial \underline{\gamma}_C}\right)^T \dot{\underline{p}}_{RC}$$
$$+ \left(\frac{\partial \underline{\gamma}_{GI}}{\partial \underline{\gamma}_C}\right)^T \dot{\underline{p}}_{GI} + \left(\frac{\partial \underline{\gamma}_{GO}}{\partial \underline{\gamma}_C}\right)^T \dot{\underline{p}}_{GO} + \left(\frac{\partial \underline{\gamma}_{L}}{\partial \underline{\gamma}_C}\right)^T \dot{\underline{p}}_L$$
$$= -A_{CI}\dot{\underline{p}}_I - A_{CRC}\dot{\underline{p}}_{RC} - A_{CGI}\dot{p}_{GI} - A_{CGO}\dot{\underline{p}}_{GO} - A_{CL}\dot{\underline{p}}_L = \dot{\underline{p}}_C \qquad (36)$$

Therefore
$$\frac{d}{dt}\left(\frac{\partial L'}{\partial \dot{\underline{\gamma}}_C}\right) - \frac{\partial L'}{\partial \underline{\gamma}_C} = \dot{\underline{p}}_C - \dot{\underline{p}}_C = 0 \qquad (37)$$

The actional functional $J(\underline{\gamma}) = \int_{t_o}^{t} L'(\underline{\gamma}_C, \dot{\underline{\gamma}}_C\ t)dt$ of the system is stationary in the fundamental TGC space, or γ_C space. Furthermore, for a simple harmonic oscillator, e.g., a mass-spring system or an LC circuit, in order to have L' satisfy the Lagrange equation for both generalized coordinates r_C and r_L, one has to assume that $\frac{\partial L'}{\partial \dot{\underline{\gamma}}_L} \equiv \underline{p}_L$. (38)

Since
$$\frac{\partial L'}{\partial \underline{\gamma}_L} = \frac{\partial(-U)}{\partial \underline{\gamma}_L} = \frac{\partial}{\partial \underline{\gamma}_L} \int_{t_o}^{t} \dot{\underline{p}}_L^T \, d\underline{r}_L = \dot{\underline{p}}_L, \qquad (39)$$
$$\frac{d}{dt}\left(\frac{\partial L'}{\partial \dot{\underline{\gamma}}_L}\right) - \frac{\partial L'}{\partial \underline{\gamma}_L} = 0 \qquad (40)$$

The action functional $J(\gamma)$ of the system is stationary both in the $\underline{\gamma}_C$ and $\underline{\gamma}_C$, $\underline{\gamma}_L$ space

The Hamiltonia of a system is defined to be

$$H \equiv - L' + \underline{p}^T \dot{\underline{\gamma}} \tag{41}$$

Therefore, $$H \equiv -(\hat{E}-U) + \underline{p}_C^T \dot{\underline{\gamma}}_C = U + \int \dot{\underline{\gamma}}_C^T \, d\underline{p}_C \tag{42}$$

Equations (31) and (42) give the new Hamiltonian energy function for the nonreciprocal dissipative system. Upon a closer observation of the expression, one can see that it includes the energy stored in the energy-storage elements, the energy dissipated in the dissipative elements both in the tree and in the cotree, and the energy supplied by the sources. For an ideal gyrator, the total energy in the input and output is zero. Therefore, the new Hamiltonian, H, is really the total energy of the system at time t.

Taking partial differentiation of H with respect to γ_C one obtains

$$\frac{\partial H}{\partial \underline{\gamma}_C} = \frac{\partial U}{\partial \underline{\gamma}_C} = -\dot{\underline{p}}_C, \tag{43}$$

$$\frac{\partial H}{\partial \underline{p}_C} = \dot{\underline{\gamma}}_C \tag{44}$$

$$\frac{\partial H}{\partial \underline{\gamma}_L} = \frac{\partial U}{\partial \underline{\gamma}_L} = \frac{\partial}{\partial \underline{\gamma}_L} \left(- \int \dot{\underline{p}}_L^T \, d\underline{r}_L \right) = - \dot{\underline{p}}_L, \tag{45}$$

and

$$\begin{aligned} \frac{\partial H}{\partial \underline{p}_L} &= \left(\frac{\partial p_C}{\partial p_L} \right)^T \dot{\underline{\gamma}}_C + \left(\frac{\partial p_V}{\partial p_L} \right)^T \dot{\underline{\gamma}}_V + \left(\frac{\partial p_{RT}}{\partial p_L} \right)^T \dot{\underline{\gamma}}_{RT} \\ &= (-A_{CL})^T \dot{\underline{\gamma}}_C + (-A_{VL})^T \dot{\underline{\gamma}}_V + (-A_{RTL})^T \dot{\underline{\gamma}}_{RT} \\ &= B_{LC} \dot{\underline{\gamma}}_C + B_{LV} \dot{\underline{\gamma}}_V + B_{LRT} \dot{\underline{\gamma}}_{RT} = -\dot{\underline{\gamma}}_L. \end{aligned} \tag{46}$$

(Examples are omitted)

IV. CONCLUSIONS

It is shown in this paper that in a topoligical generalized coordinates [TGC] space and its fundamental TGC space every lumped bilateral physical system is conservative...when all components are included...and the corresponding action functional is stationary.

The energy associated with dissipative components should be included in the potential energy function. The Hamiltonian function thus derived is the total energy of the system and is constant if H is not an explicit function of time. Furthermore, H satisfies the canonical-form Euler-Lagrange equations, no matter whether the system is linear, nonlinear, reciprocal, nonreciprocal, and dissipative or nondissipative.

"And if any man think that he knoweth anything, he knoweth nothing yet as he ought to know." ------I Corinthians 8:2

"For I determined not to know any thing among you, except Jesus Christ, and Him Crucified." -----I Corinthians 2:2

ACKNOWLEDGEMENT

The author wishes to thank Dr. W. W. Cannon, Chairman, Department of Electrical Engineering, West Virginia University for his encouragement and Mrs. Lynette Cross for her skillfully typing the manuscript.

BIBLIOGRAPHY

1. Koenig, and Blackwell, "On the Codification of Lagrangian Formulation", Proc. IRE, 1958.
2. Blackwell, W. A., "Hamiltonian Approach to Conservative Networks with Ideal Sources" (unpublished).
3. White, H. J., "Role of the Generalized Methods of Mechanics in Elec. Eng.", IEEE Trans. on Ed., Mar. 1968.
4. Rogers, J. W., "Finding Generalized Forces by Use of Partial Derivatives of Energy Functions", IEEE Trans. of Ed., June 1967.

5. Goldstein, H., Classical Mechanics. Addison-Wesley Pub. Co., 1959.
6. Lindsay and Margenau, Foundations of Physics, John Wiley & Sons, 1944.
7. Margenau and Murphy, The Mathematics of Physics and Chemistry, D. Van Nostrand Co., 1950.
8. Slater and Frank, Introduction to Phsyics, McGraw-Hill, 1933.
9. Wells, D. A., Theory and Problems of Lagrangian Dynamics, Schaum Pub., 1967.
10. Zadeh, L. A. and E. Polak, System Theory, McGraw-Hill, 1969.
11. Gossick, B. R., Hamilton's Principle and Physical Systems, Academic Press, 1967.
12. Koenig and Blackwell, Electromechanical System Theory, McGraw-Hill, 1961.
13. Blackwell, W. A., Mathematical Modeling of Physical Networks, MacMillan 1968.
14. Gelfand and Fomin, Calculus of Variations, Prentice-Hall, 1963.
15. Blackwell and Grigsby, "A Simple Formulation Procedure for State-Space Models of Electric Networks", J. of the Franklin Institute, June 1966.
16. Koenig, Tokad, and Kesavan, Analysis of Discrete Physical Systems, McGraw-Hill, 1967.
17. Bryant and Bers, "The Degrees of Freedom in RLC Networks", IRE Trans. on C.T., June 1960.
18. Roe, P. H., Networks and Systems, Addison-Wesley Publishing Co., 1966.
19. Trent, H. M., "Isomorphisms between Oriented Linear Graphs and Lumped Physical Systems", J. of the Acoustical Soc. of America, Vol. 27, May 1955.
20. Firestone, F. A., "A New Analogy between Mechanical and Electrical Systems", J. of the Acous. Soc. of America, Vol. 4, Jan. 1933.
21. Cheng, D. K., Analysis of Linear Systems, Addison-Wesley, 1961.
22. Martens and Allen, Introduction to System Theory, Charles E. Merrill, 1969.
23. Meisel, J., "State Equations of Electrical Networks from a Lagrangian Formulation", IEEE Trans. on C.T., May 1969.
24. Spiegel, M. R., Theory and Problems of Theoretical Mechanics with an Introduction to Lagrange's Equations and Hamiltonian Theory. Schaum Publ. Co., 1967.
25. Stern, T.E., Theory of Nonlinear Networks and Systems, Addison-Wesley 1965.
26. Huang, "State-space Models of Lumped Linear Bilateral Physical Systems by Hamiltonian/Linear Graph Approach", Ph.D. Dissertation, VPI 1969.
27. Huang and Blackwell, "New Lagrangian and Hamiltonian Functions for Linear Dissipative Physical Systems with Ideal Drivers", IEEE Trans. on C.T., Vol. CT-18, July 1971.
28. White and Woodson, Electromechanical Energy Conversion, John Wiley, 1959.
29. Synge and Griffith, Principle of Mechanics, McGraw-Hill, 1949.
30. Seely and Goldstein, Jr., "State Function Ajproach to System Analysis", J. Franklin Inst., Vol. 283, March 1967.
31. Brayton and Moser, "A Theory of Nonlinear Networks, I" and "A Theory of Nonlinear Networks, II", Q. Appl. Math. Vol. XXII, Apr. & Jul. 1964.
32. Evans, Jones, and Sen, "Canonical Formulation of Variationsl Principles in Nonconservative Networks", Proc. 13th MWSC, Univ. of Min., May 1970.
33. Sen and Evans, "Formulation of Generalized Networks in Terms of Hamiltonian Energy Function", Electronics Letters, Vol. 6, 1970.
34. Chua and McPherson, "Explicit Topological Formulation of Lagrangian and Hamiltonian Equations for Nonlinear Networks", Proc. 1973 IEEE ISCT.
35. Biorci, G., Network and Switching Theory, Academic Press, 1968.
36. Saletan, E. J. and A. H. Cromer, Theoretical Mechanics, Wiley 1971.
37. Murdoch, J. B., Network Theory, McGraw-Hill, 1970.
38. Barger, V. and Olsson, M., Classical Mechanics, McGraw-Hill, 1973.
39. Mizushima, Masataka, Theoretical Physics, John Wiley & Sons, 1972.

THE ANALYSIS OF THE BORDER VIOLATION PROBLEM AS A POINT PROCESS GAME

ANTHONY EPHREMIDES
Electrical Engineering Department
University of Maryland
College Park, Maryland 20742

Abstract

A new class of differential games is considered, where the players are controlling random point processes. Such games can be modeled mathematically in a standard fashion by state and observation equations; however they present substantial conceptual and technical differences from other stochastic games. An example of a point process game, the border violation problem, is analyzed. Complete solutions are derived for simple forms of this problem, and procedures and partial results are obtained for its extended versions.

Introduction

From the study of ordinary queueing systems, like for example the classical Gl/M/S, in their natural competitive or co-operative context, one is led to the formulation of their design as a differential game problem. It can then be observed that this formulation covers a number of other problems as well. For example border patrolling problems, traffic control situations, surveillance radar systems, and inventory analysis may all be described by similar mathematical models. These models involve essentially a vector state equation and a vector observation equation where the control functions to be chosen by the players of the game are,or look like, intensity functions of random point processes. The "state" of such a system is a vector of plain probabilities, each representing the likelihood of a "server" being busy or occupied as a function of time. The observation process is just a random point process of "arrivals" or occurrences. Already one can identify some features of this model that distinguish it from similar ones in differential game theory.

First, the state is not directly observable. As a matter of fact the way it is imbedded in the observation equation may be complicated. Secondly, the game is stochastic, but not in the additive noise fashion as most stochastic game models [1, 2] studied so far are. Thirdly, the usual complications of constraints imposed on the state as well as on the controls are present.

Of course there are some redeeming features in these point process games. Namely the state equation is linear and the controls are separable. Thus a good portion of existing theory can be used. However the system is bilinear, in that the controls multiply the state, and therefore its analysis is far from straight-forward. Finally even in the fairly general versions

---

This work has been supported by NSF Grant GK32709.

of these games simple linear cost functions enter as natural ones, although one can consider a variety of other reasonable criteria. Besides, in almost all cases the game involves two players and is zero-sum.

The concept of a point process game can be best described in terms of a typical example that exhibits all the properties discussed so far. In this paper, the border violation problem has been chosen for this purpose.

The Border Violation Problem

Description: A border can be penetrated through several corridors each of which is protected by a number of sentries. In a typical corridor it is assumed that attackers are dispatched randomly in time with

$$p\{\text{one arrival in } [t, t+dt)\} = \alpha_t \, dt$$

The quantity α_t is the intensity rate of the offensive point process which is assumed to be Poisson. Of course one may consider only a regular [3-5] process by defining $\alpha_t dt$ as the conditional probability given the σ-field of the past. A single sentry is available to protect the corridor. It is assumed that he engages the attackers for a period of time γ_s, starting at $t=s$. At the end of the engagement (fight) period the attacker is neutralized. The "service" time γ_s is a random variable, independent of the attack process, and distributed in an exponential-like fashion. Namely, at any instant $t > s$

$$p\{\gamma_s \leq t-s+dt / \gamma_s > t-s\} = \beta_t dt$$

Thus β_t can be thought of as the defense capability; generally, the larger its value, the smaller the average service time. Realistically one may interpret the liquidation of a sentry as followed by an instant replacement, hence his invincibility. Therefore an intruder may penetrate only if at the time of his arrival the sentry is busy.

The objective of the attack strategist, termed player 1, is to maximize the average number of penetrators over a fixed period of time T, whereas the objective of the defender, termed player 2, is to minimize the same quantity. Player 1 is free to choose $\alpha_t, 0 \leq t \leq T$, and player 2 selects β_t, $0 \leq t \leq T$. Thus they may be called the α-player and β-player respectively. As seen later one may approximate a real conflict by allowing player 1 to choose the actual instants of attack, rather than their density and player 2 to control more than just his "average" service time.

Of course some cost weight must be placed upon the choices of α_t, and β_t. Thus, if M_t is the counting process of penetrators, the cost function of this zero-sum, two-player game is defined by

$$K = E[M_T] + c \cdot \int_0^T (\beta_t - \alpha_t) dt \qquad (1)$$

Under the previous assumptions it can be easily shown that

$$E[M_T] = \int_0^T \alpha_t \, p_t \, dt$$

where p_t is the probability that the sentry is engaged at time t, and that

$$\dot{p}_t = -(\alpha_t + \beta_t)p_t + \alpha_t \tag{2}$$

with the choise of initial condition given by

$$p_0 = 0$$

By regarding p_t as the "state" of the system, Eq. (2) is the system equation. The observation equation is given by

$$dy_t = dN_t \tag{2a}$$

where N_t is the attack counting process (Poisson) or by

$$dy_t = dM_t \tag{2b}$$

where M_t is the penetrator counting process (Poisson too, under some conditions). For a first analysis one may consider

$$y_t = p_t$$

that is, perfect information on the state.

Comments: 1) In view of the fact that the system is linear with a linear cost, solutions on the boundary are expected. Hence a reasonable set of constraints for the control functions α_t, β_t is that they be bounded. Thus

$$0 \le \alpha_t \le \alpha_M \quad , \quad 0 \le t \le T$$
$$0 \le \beta_t \le \beta_M \quad , \quad 0 \le t \le T$$

Furthermore we will look for piece-wise continuous rather than measurable functions. 2) The necessary constraint for the state is that

$$0 \le p_t \le 1 \quad , \quad 0 \le t \le T$$

which defines the "playing space" $[0,1] \times [0,T]$. This is automatically ensured by the form of Eq. (2) and by the chosen initial condition. 3) The general form of the chosen cost function allowing for randomized strategies of strategies that are functions of the random observation process y_t (or its entire past) is given by

$$K(\alpha_t, \beta_t, p_t) = E\left\{\int_0^T \alpha_t \, p_t \, dt + c\int_0^T (\beta_t - \alpha_t)dt\right\}$$

where

$$\alpha_t = \alpha'(y_s, s \le t) = \alpha(p_s, s \le t \, ; \omega) \tag{2c}$$

$$\beta_t = \beta'(y_s, s \le t) = \beta(p_s, s \le t \, ; \omega)$$

4) Since the system is linear, a quadratic cost function might facilitate the analysis. With the system described above it is not easy to formulate a quadratic criterion. For example an attempt to consider as cost the

variance of N_T given that $E[N_T] <$ threshold, yields an even more complicated criterion function. However possibilities for other criteria include requiring $p_t \gtrless$ threshold. The bilinear structure of the system and the availability of results [6,7] in the controllability of bilinear systems suggest such an approach. That might be an appropriate criterion in the surveillance radar system case or in other communication-detection versions of point process games.

Perfect Information Case

In this case it is assumed that both players can observe the state p_t at every $0 \le t \le T$. First, one verifies that the system satisfies trivially various sets of sufficient conditions for the existence of a saddle point, as for example the ones in Thm. 2.5.1 of [8], namely the existence of a unique solution to the state equation for any given controls α_t, β_t, the latter's continuity in the controls, and the linearity of the pay-off in α_t, β_t. That the saddle point consists of pure strategies will be concluded a posteriori, after finding and exhibiting such pure strategies. To find them, Pontryagin's maximum (minimum) principle will be used [9] in the standard control theoretic fashion by finding the $\max_\alpha[\min_\beta(K)]$, which, by the existence of value for the game, will be equal to $\min_\beta[\max_\alpha(K)]$. This approach has the advantage of supplying the solution to the control problem of designing the defense β_t for known rate α_t (typical in simple models of air traffic control systems), by specifying the function $\beta_t(\alpha_t)$ in the intermediate stage of the maxmin procedure.

Of course the analysis can be carried out in terms of cost-to-go from an arbitrary starting position (π, τ) in the playing space $[0,1] \times [0, T]$. Thus prescriptions for optimal strategies will be supplied to each player in case the opponent deviates from optimal play and consequently forces the state out of the optimal trajectory. Furthermore arbitrary boundary conditions for p at the beginning of the play must be allowed.

The Min part: The system is described by

$$\dot{p}_t = -(\alpha_t + \beta_t)p_t + \alpha_t$$
$$\dot{q}_t = (\alpha_t + \beta_t)q_t - \alpha_t$$

where q_t is the co-state, and with boundary conditions

$$p_\tau = \pi \quad , \quad q_T = 0$$

The Hamiltonian is given by

$$H(p_t, q_t; \alpha_t, \beta_t) = (c - p_t q_t)\beta_t + (p_t + q_t - p_t q_t - c)\alpha_t$$

The H-minimal control will be

$$\beta_t = \begin{cases} 0 & , \; p_t q_t < c \\ \text{indeterminate} & , \; p_t q_t = c \\ \beta_M & , \; p_t q_t > c \end{cases}$$

Already it should be noted that for $c \ge 1$ the overall game assumes the trivial solution $\alpha_t \equiv \beta_t \equiv 0$, since it is then more expensive to both players to play the game. So it will be assumed that $0<c<1$.

The solution is illustrated in figures 1 and 2. As it turns out the relative magnitude of β_M and α_t is critical to the form of the solution. In figure 1 the case of the so called weak defense is depicted, that is when

$$\beta_M \leq \min_{0\leq t\leq t_1} \alpha_t \frac{1-\sqrt{c}}{\sqrt{c}}$$

In figure 2 we have the case of strong defense, that is

$$\beta_M \geq \max_{0\leq t\leq t_1} \alpha_t \frac{1-\sqrt{c}}{\sqrt{c}}$$

For the case that

$$\min_{0\leq t\leq t_1} \alpha_t \frac{1-\sqrt{c}}{\sqrt{c}} < \beta_M < \max_{0\leq t\leq t_1} \alpha_t \frac{1-\sqrt{c}}{\sqrt{c}}$$

which represents an intermediate situation we get the answer in all but a portion of the playing space, as will be explained later. However in the maximization portion of the maxmin process this difficulty will be circumvented. The quantities that appear in the figures that need explanation are the following. The two critical instants t_1, and t_2 are defined by

$$\int_{t_1}^{T} \alpha_t dt = \ell n \frac{1}{1-\sqrt{c}} \tag{3}$$

$$\int_{t_2}^{T} \alpha_t dt = \ell n \frac{1}{1-c} \tag{4}$$

Eq. (3) implies that $q_{t_1} = \sqrt{c}$, and Eq. (4) that $q_{t_2}=c$ while the function ψ_t denotes the locus of the points at which $p_t q_t = c$, with $\psi_t = p_t$.

In fig. 1 the function θ_t is defined parametrically by

$$c=\left\{\psi_s - \int_t^s \alpha_\xi \exp[-\int_\xi^s(\alpha_\eta+\beta_M)d\eta]d\xi\right\}\left\{\frac{c}{\psi_s}+\int_t^s \alpha_\xi \exp[\int_\xi^s (\alpha_\eta+\beta_M)d\eta]d\xi\right\} \tag{5}$$

and

$$\theta_t = \psi_s - \int_t^s \alpha_\xi \exp[-\int_\xi^s (\alpha_\eta+\beta_M)d\eta]\, d\xi\ , \quad t_1\leq s<t_2\ , \ t\leq t_1$$

In fig. 2 this function does not show because the above equations do not have solutions. The fact that we do not know the conditions for existence of solutions in the intermediate case accounts for the incomplete specification of the solution as shown in fig. 3. The function θ_t represents the locus of the points at which $p_t \cdot q_t = c$ with $\theta_t = p_t$.

The Max part: There are three regimes of solutions for β_t in the min part, namely $\beta=0$, β=indeterminate, and $\beta=\beta_M$. For each regime Pontryagin's principle is used to obtain that the H-maximal control is given by

$$\alpha_t = \begin{cases} 0 & , \quad p_t+q_t-p_tq_t< c \\ \text{indeterminate} & , \quad p_t+q_t-p_tq_t=c \\ \alpha_M & , \quad p_t+q_t-p_tq_t> c \end{cases}$$

for the regimes corresponding to $\beta=0$ and $\beta=\beta_M$, and by

$$\alpha_t = \begin{cases} 0 & , \quad p_t+q_t-(1/\sqrt{c})p_tq_t < c \\ \text{indeterminate} & , \quad p_t+q_t-(1/\sqrt{c})p_tq_t = c \\ \alpha_M & , \quad p_t+q_t-(1/\sqrt{c})p_tq_t > c \end{cases}$$

for the regime β=indeterminate. In all cases it is argued that α_t cannot be indeterminate. Therefore always α_t is either α_M or 0. The solution is illustrated in figures 4 and 5, for the cases of weak, and strong defenses respectively. The intermediate case does not arise, since in the region of possible ambiguity we have $\alpha_t = \alpha_M$, so that the defense can be determined to be either weak or strong. The function δ_t, which is a dispersal surface, as one can easily show by computing the pay-offs, and below which we have the only region where $\alpha_t = 0$, is defined by

$$\delta_t = 1 - \frac{\alpha_M(1-c)(T-t)}{1-\exp[-\alpha_M(T-t)]}$$

Thus the closed-loop, feedback strategies are completely specified for both players, in the perfect information case. The solution to equation(5) is given by

$$t = s - \frac{1}{\alpha_M+\beta_M}\, \ell n \frac{\dfrac{\alpha_M}{\alpha_M+\beta_M} - \dfrac{c}{\psi_s}}{\dfrac{\alpha_M}{\alpha_M+\beta_M} - \psi_s}\ , \quad t_1 \le s \le t_c < t_2$$

when

$$\beta_M < \alpha_M \frac{1-\sqrt{c}}{\sqrt{c}} .$$

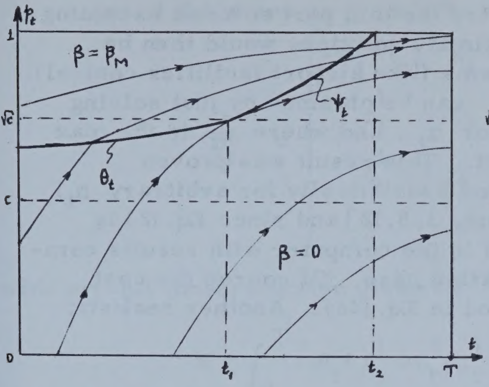
Imperfect Information Case

The quantity p_t that was considered to be the "state" in the formulation of the problem is clearly unobservable and one can only compute it from the differential equation (2) if α_t and β_t are completely known. This not being the case it is clear that the observation equation (2a) or (2b) has to be considered. Since a solution for the perfect information case has been obtained, one would hope that a separation theorem [10, 11] here might hold; that is, that Eqs. (2c) might take the form

$$\alpha_t = \alpha(\hat{p}_t) \quad , \quad \beta_t = \beta(\hat{p}_t)$$

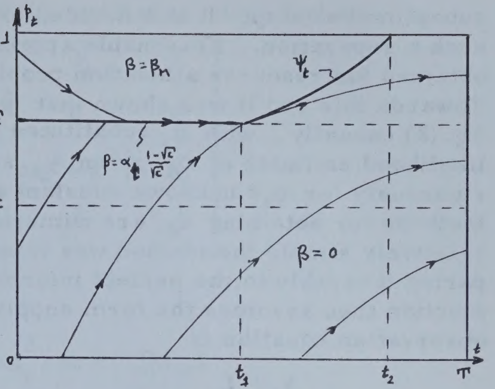
where $\alpha(\cdot)$, $\beta(\cdot)$ are the saddle point solution derived before, and $\hat{p}_t$ the max likelihood estimate of p_t given y_s, $s\le t$.

Showing a separation theorem has not been accomplished. Thus, as a



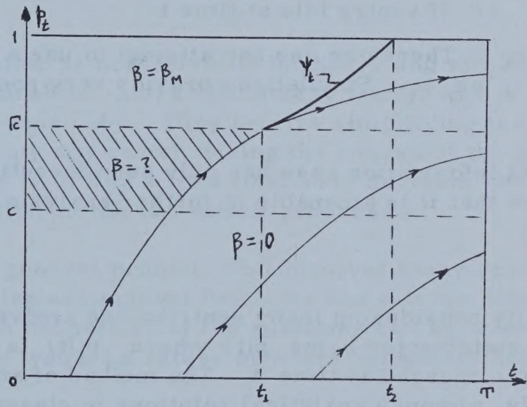
p_t
1
$\beta = \beta_M$
$\sqrt{c}$
ψ_t
θ_t
c
$\beta = 0$
0
t_1
t_2
T
t

Fig. 1: H-minimal β_t for given α_t and for weak defense



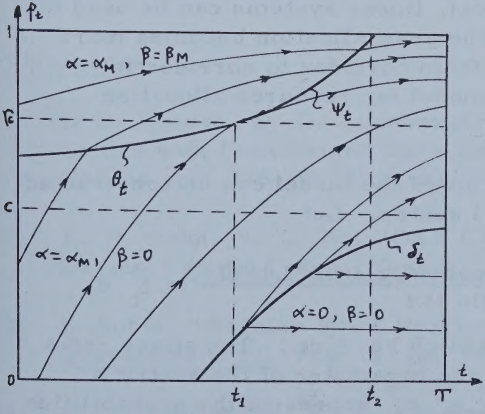
p_t
1
$\beta = \beta_M$
ψ_t
$\sqrt{c}$
$\beta = \alpha_t \frac{1-\sqrt{c}}{\sqrt{c}}$
c
$\beta = 0$
0
t_1
t_2
T
t

Fig. 2: H-minimal β_t for given α_t and for strong defense.



p_t
1
$\beta = \beta_M$
ψ_t
$\sqrt{c}$
$\beta = ?$
c
$\beta = 0$
0
t_1
t_2
T
t

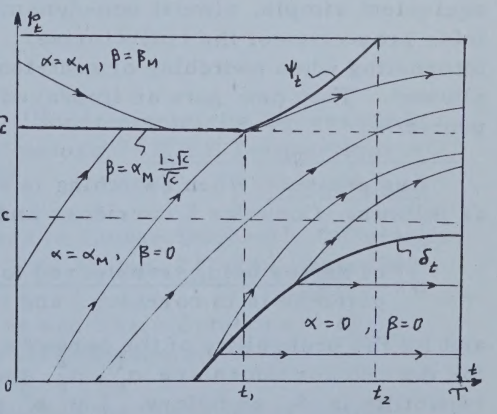
Fig. 3: H-minimal β_t for given α_t and for intermediate defense.



p_t
1
$\alpha = \alpha_M$, $\beta = \beta_M$
$\sqrt{c}$
ψ_t
θ_t
c
$\alpha = \alpha_M$, $\beta = 0$
δ_t
$\alpha = 0$, $\beta = 0$
0
t_1
t_2
T
t

Fig. 4: Saddle point solution regimes for weak defense

$\beta_M < \alpha_M \frac{1-\sqrt{c}}{\sqrt{c}}$



p_t
$\alpha = \alpha_M$, $\beta = \beta_M$
ψ_t
$\sqrt{c}$
$\beta = \alpha_M \frac{1-\sqrt{c}}{\sqrt{c}}$
c
$\alpha = \alpha_M$, $\beta = 0$
δ_t
$\alpha = 0$, $\beta = 0$
0
t_1
t_2
T
t

Fig. 5: Saddle point solution regimes for strong defense

$\left(\beta_M \geq \alpha_M \frac{1-\sqrt{c}}{\sqrt{c}}\right)$

suboptimal attempt, it was decided to try the min part solution assuming such a separation. Reasonable approximate solutions would then be obtained for resource allocation problems (like airport facilities control). Towards this end it was shown that $\hat{p}_t$ can be obtained by just solving Eq. (2) causally, with $\hat{\alpha}_t$ substituted for α_t, and where $\hat{\alpha}_t$ is the max likelihood estimate of α_t given $y_s, s \leq t$. This result was proven rigorously for $\alpha_t \equiv$ unknown constant and heuristically for arbitrary α_t. Methods for obtaining $\hat{\alpha}_t$ are numerous [3, 5, 12] and since Eq. (2) is relatively simple the method was tried in the computer with results comparing favorably to the perfect information case. Of course the cost function then assumes the form supplied in Eq. (2a). Another realistic observation equation is

$$y_t = I_t$$

where I_t is an indicator function of engagement of the sentry, given by

$$I_t = \begin{cases} 1 & \text{if sentry busy at time } t \\ 0 & \text{if sentry idle at time } t \end{cases}$$

Clearly $E[I_t] = p_t$. Therefore one can attempt to use a separation procedure using I_t as $\hat{p}_t$. Simulations provide very good results in this case.

The imperfect information case has only been investigated imperfectly, but indications are that it is amenable to further analysis.

Generalizations

Dimension: By considering many sentries one arrives at a matrix equation with the state vector being $\overline{p}(t)$ where $p_i(t)$ is the probability that i sentries are engaged at time t. The method of analysis generalizes straight-forwardly, although analytical solutions in closed form are hard to obtain. Similar extensions can be accomplished when more than one corridor is considered. The form of the system is then nontrivially bilinear and recent results [13], in studying such systems through their equivalent simple, almost non-dynamical, linear systems can be used to infer properties of the trajectories. The generalization becomes more interesting when switching of sentries from corridor to corridor is allowed. Then one gets an improved model for resource allocation problems.

Assignment: When switching is allowed the model can be constructed as follows. Consider 2 corridors and 1 sentry. Let

$$p\left[\frac{\text{of sentry being transferred to corridor 2 in } [t, t+dt)}{\text{given he is in corridor 1 and idle at t}}\right] = \delta_t^1 \; dt$$

and let the probability of the opposite switch be $\delta_t^2 dt$. The attack rates for the two corridors are α_t^1, α_t^2 and the parameter of the sentry's capability is β_t as before. Let $p_t^0, p_t^1, p_t^2, p_t^3$ represent the probabilities of the four possibilities that the sentry is in corridor 1 and idle, or in 1 and busy, or in 2 and idle, or in 2 and busy at time t respectively. Then allowing for switches only during idle periods and denoting by $\overline{p}_t$ the

the vector of the previous probabilities the system is described by

$$\dot{\bar{p}}_t = \begin{bmatrix} -(\alpha_t^1+\delta_t^1) & \beta_t & \delta_t^2 & 0 \\ \alpha_t^1 & -\beta_t & 0 & 0 \\ \delta_t^1 & 0 & -(\alpha_t^2+\delta_t^2) & \beta_t \\ 0 & 0 & \alpha_t^2 & -\beta_t \end{bmatrix} \bar{p}_t$$

with pay-off (cost) function given by

$$K = \int_0^T c(\beta_t + \delta_t^1 + \delta_t^2 - \alpha_t^1 - \alpha_t^2)\, dt + E[M_T]$$

It is of course assumed that the two attack processes are mutually independent and that the switching intensities are independent of the service times.

The above system is non-trivially bilinear, and similar analysis as before is still applicable. More elementary treatments of restricted cases have been already done [14]. They involve simplifying the assignment strategy to allow only one switch during the course of the game. Then the optimum delta rule is to find a max likelihood decision time t_c and then make a MAP decision on the two attack processes.

Of course the general problem that involves many corridors, several sentries and includes assignment functions has a large dimensionality and therefore, although properties of the solutions can be infered analytically, the actual solutions would be rather obtained computationally.

Finally the possibility of sentry liquidation can be incorporated in the previous model by defining an additional intensity h_t such that $h_t\, dt$ represents the probability of elimination in $[t, t+dt)$ conditioned appropriately on the sentry' s tate at time t.

References

1. I. B. Rhodes, D. G. Luenberger, "Nondeterministic Differential Games with Constrained State Estimators," IEEE Transactions AC, Vol. 14, pp. 476-481, October 1969.

2. R. D. Behn, Y. C. Ho, "On a Class of Linear Stochastic Differential Games," IEEE Transactions AC, Vol. 13, pp. 227-240, June 1969.

3. I. Rubin, "Regular Point Processes and Their Detection," IEEE Transactions IT, Vol. 18, pp. 547-557, September 1972.

4. F. J. Beutler, O. A. Z. Leneman, "On the Statistics of Random Pulse Processes," Information and Control, Vol. 18, pp. 326-341, 1971.

5. D. L. Snyder, "Filtering and Detection for Doubly Stochastic Poisson Processes," IEEE Transactions IT, Vol. 18, pp. 91-102, January 1972.

6. R. W. Brockett, "System Theory on Group Manifolds and Coset Spaces," SIAM Journal on Control, Vol. 10, p. 265, May 1972.

7. J. Wei, E. Norman, "Lie Algebraic Solution of Linear Differential Equations," Journal of Mathematical Physics, Vol. 4, pp. 575-581, 1963.

8. A. Friedman, Differential Games, Wiley-Interscience, 1971.

9. M. Athans, P. Falb, Optimal Control, McGraw-Hill, 1966.

10. W. M. Wonham, "On the Separation Theorem of Stochastic Control," SIAM Journal on Control, Vol. 6, pp. 312-326, 1968.

11. D. L. Snyder, I. B. Rhodes, "Filtering and Control Performance Bounds with Implications on Asymptotic Separation," Automatica, Vol. 8, pp. 747-753, 1972.

12. D. L. Snyder, "Smoothing for Doubly Stochastic Poisson Processes," IEEE Transactions IT, Vol. 18, pp. 558-562, September 1972.

13. J. T. Lo, Private Communication, September 1973.

14. R. J. Corn, A. Ephremides, "Point Process Games," Proceedings 1972 IEEE Conference on Decision and Control, p. 198, December 1972.

SMALL SAMPLE ESTIMATION FROM INFORMATION THEORETIC METHODS[1]

R. L. KASHYAP
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

Given an observation Z obeying a binomial distribution, we will determine an estimate of the unknown parameter λ in the distribution, the so-called divergence estimate, by means of the Bayes approach using a particular prior cumulative distribution function (CDF), F*. This CDF is determined by means of a two-person, zero-sum game associated with the estimation problem using the expected value of the divergence function as the payoff function for the game. This optimal CDF also maximizes the uncertainty about the occurence of the observation Z.

When the sample size N is large, the mean square error of the divergence estimate is nearly equal to the minimum possible value. When the sample size N is small, we will show, by means of examples, that the variance of the estimate is much smaller than that of traditional estimates like maximum likelihood, over a wide range of the unknown parameter.

A good approximation to the optimal CDF F* is easily computed by recognizing that the CDF approximates a uniformly spaced staircase function

I. INTRODUCTION

We will consider the minimax estimation of the parameters in a binomial distribution using a particular non quadratic loss function, the so-called divergence loss function. Our initial motivation for such a study was only curiosity. Almost all the minimax estimation problems treated in statistical literature [1] use the quadratic loss function. It is interesting to find how the structure of the minimax solution obtained from the divergence loss function differs from that with the quadratic loss function.

A second reason is that the minimax estimate has some appealing features in the class of Bayes estimate, since an ordinary Bayes estimate is computed from an arbitrary prior probability distribution whereas the prior probability distribution for the minimax estimate has been chosen so that the maximum possible risk associated with it has the smallest possible value. But the minimax estimate of the binomial distribution computed from the quadratic loss function is not entirely satisfactory as shown in [2], since the estimate is not asymptotically efficient. The solution given here using the divergence loss function does not suffer from such difficiencies.

Moreover we will discuss the circumstances under which the different estimates like the minimax estimate mentioned here, the maximum likelihood estimate, the Bayes estimate with uniform prior distribution etc. may be appropriate to use, especially in the case of small samples.

Finally, the minimax estimate discussed here has some interesting information theoretic interpretations, since it is derived from the divergence loss function.

---

[1] This work was partially supported by NSF under Grant GK-18225 and AFOSR under Grant 69-1776.

II. THE TWO PERSON ZERO SUM GAME

Consider the binomial distribution in (2.1)

$$P(z|\lambda,N) = \binom{N}{z} \lambda^z (1-\lambda)^{N-z}, \quad 0 \leq \lambda \leq 1 \tag{2.1}$$

Let a be an estimate of λ, $0 < a < 1$

Loss function:

The loss function L, the so called divergence loss function, will be used to measure the discrepancy between a and λ,

$$L(\lambda,a) = \lambda \ln \frac{\lambda}{a} + (1-\lambda) \ln \frac{1-\lambda}{1-a}$$

Note that

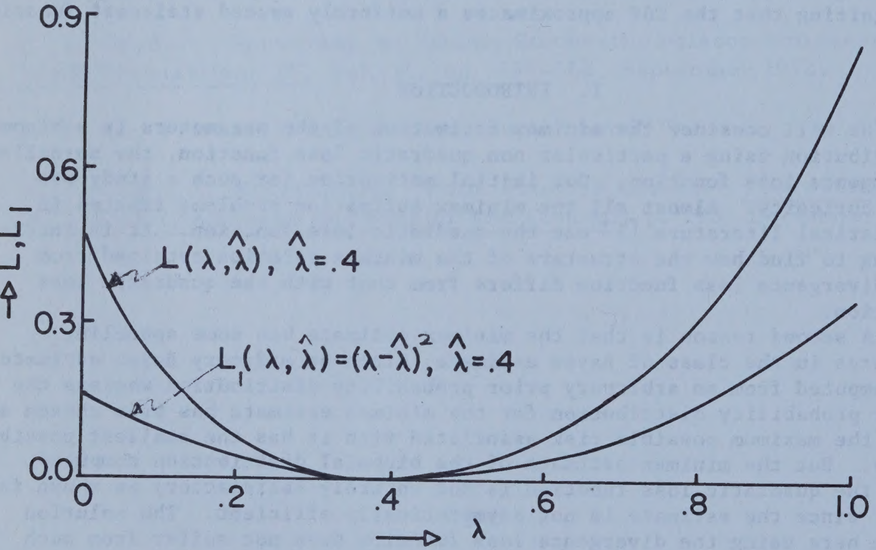
$$\begin{aligned} L(\lambda,a) &> 0 \text{ if } \lambda \neq a, \text{ since } \ln x \leq x - 1 \\ &= 0 \text{ if } \lambda = a \end{aligned}$$

$$L(\lambda,a) \neq L(a,\lambda)$$

The loss function L(.) given above is compared graphically with the usual quadratic loss function L_1 in figure 1

$$L_1(\lambda,a) = (\lambda - a)^2$$

The divergence loss function L penalizes the larger errors more than the quadratic function.



Payoff function:

Let $Z = \{0,1,\ldots,N\}$. Let b(z) be an estimate of λ based on z, i.e., b is a map defined below

$$b: Z \to (0,1), \ b \in \mathcal{B}$$

The letter b by itself will denote the map or function whereas b(z) will denote the value of the function for the argument z.

r(λ,b) = the risk function for the estimate b using the loss function L

$$= \sum_{z=0}^{N} L(\lambda,b(z))P(z|\lambda) \tag{2.3}$$

Let $F(.)$ be a prior cumulative distribution function of λ, $F \in \mathcal{F} \triangleq$ set of all possible CDF's

$R(F,b)$ = Average risk function using the prior CDF
$= \int r(\lambda,b)dF(\lambda)$

Consider the two person zero sum $(\mathcal{F},\mathcal{B},R)$ where one player chooses a $b \in \mathcal{B}$ to minimize R whereas the other player (nature) chooses a $F \in \mathcal{F}$ to maximize R.

This game has a saddle point in pure strategies and is given in the following theorem.

Notation: b_F = the Bayesian estimate of λ with the prior C.D.F, F

Theorem 1: Consider the game $(\mathcal{F},\mathcal{B},R)$

$$\text{Let } F^* = \text{Arg.}\left\{\sup_{F\in\mathcal{F}} \psi(F)\right\}, \quad \psi(F) = R(F,b_F) \tag{2.4}$$

and $b^* = b_{F^*}$

Then the pair (F^*,b^*) obeys the following saddle point inequality (2.5)

$$R(F,b^*) \le R(F^*,b^*) \le R(F^*,b) \tag{2.5}$$

Proof of Theorem 1:

The existence of F^* will be demonstrated later.

$$\begin{aligned} R(F,b) &= \sum_{z=0}^{N} \int dF(\lambda)P(z|\lambda)\left[\lambda \ln \frac{\lambda}{b(z)} + (1-\lambda)\ln\frac{1-\lambda}{1-b(z)}\right] \\ &= \int dF(\lambda)[\lambda \ln\lambda + (1-\lambda)\ln(1-\lambda)] \\ &\quad - \sum_{z=0}^{N}\int dF(\lambda)P(z|\lambda)[\lambda \ln b(z) + (1-\lambda)\ln(1-b(z))] \end{aligned} \tag{2.7}$$

Similarly,

$$R(F,b_F) = -\int dF(\lambda)H(\lambda) - \sum_{z=0}^{N} dF(\lambda)P(z|\lambda)[\lambda \ln b_F(z) + (1-\lambda)\ln(1-b_F(z))] \tag{2.8}$$

Subtract (2.7) from (2.8)

$$\begin{aligned} R(F,b_F) - R(F,b) &= \sum_{z=0}^{N} \underbrace{\int dF(\lambda)P(z|\lambda)}_{= \, dF(\lambda|z)P(z)}\left[\lambda \ln\frac{b(z)}{b_F(z)} + (1-\lambda)\ln\frac{1-b(z)}{1-b_F(z)}\right] \\ &= \sum_{z=0}^{N} P(z)\left[b_F(z)\ln\frac{b(z)}{b_F(z)} + (1-b_F(z)\ln\frac{1-b(z)}{1-b_F(z)}\right] \end{aligned}$$

since $\int \lambda dF(\lambda|z) = b_F(z)$

$$\le 0, \text{using the inequality } \ln x \le x - 1 \tag{2.9}$$

(2.9) is true for all F, including F = F*.

Hence, $R(F^*,b^*) - R(F^*,b) \le 0$ (2.10)

This proves the right hand side of the inequality (2.5).

To prove the left hand side of (2.5) note that F* is the value of F which maximizes the functional $\psi(F)$ subject to the constraints (2.11)

$$\left.\begin{aligned} &\int dF(\lambda) = 1 \\ &dF(\lambda) \ge 0 \;\; \forall\, \lambda \in [0,1] \end{aligned}\right. \tag{2.11}$$

Consider the augmented functional $\bar{\psi}$ defined below:

$$\bar{\psi}\,(F) = \psi\,(F) + \mu\left(\int dF(\lambda) - 1\right) + \int c(\lambda)\,dF(\lambda) \tag{2.12}$$

By the theory of Lagrange multipliers [4] there exists a scalar multiplier μ and a function c(.) obeying (2.13) such that (2.14) is valid.

$$\left.\begin{aligned} &c(\lambda) \geqslant 0 \\ &\int c(\lambda)\,dF^*(\lambda) = 0 \end{aligned}\right\} \tag{2.13}$$

$$\bar{\psi}\,(F^*) \ge \bar{\psi}\,(F) \qquad \forall\, F \subset \mathcal{F} \tag{2.14}$$

Equation (2.14) implies the eqn. (2.15)

$$\delta\bar{\psi}\,(F) = \lim_{\alpha \to 0} \frac{\partial \psi\,(F^* + \alpha G)}{\partial \alpha} \le 0 \qquad \forall\, G \tag{2.15}$$

$$\delta\bar{\psi} = \lim_{\alpha \to 0} \frac{\partial}{\partial \alpha}\Big\{R(F^* + \alpha\, G, b_{F^*+\alpha G}) + \mu\Big[\int (dF^*(\lambda) + \alpha\, dG(\lambda)) - 1\Big] + \int c(\lambda)\,dF^*(\lambda) + \alpha\, dG(\lambda)\Big\} \tag{2.16}$$

$$\lim_{\alpha \to 0} \frac{\partial}{\partial \alpha} R(F^* + \alpha G\,, b_{F^* + \alpha G}) = \lim_{\alpha \to 0} \frac{\partial}{\partial \alpha} \int (dF^*(\lambda) + \alpha\, dG\,(\lambda))\, r(\lambda, b_{F^* + \alpha G})$$

$$= \int dG(\lambda)\, r(\lambda, b_{F^*}) + \int dF^*(\lambda) \lim_{\alpha \to 0} \frac{\partial}{\partial \alpha} (r(\lambda, b_{F^* + \alpha G}))$$

$$\lim_{\alpha \to 0} \frac{\partial}{\partial \alpha} \int dF^*(\lambda)\, r(\lambda, b_{F^* + \alpha G}) = \lim_{\alpha \to 0} \frac{\partial}{\partial \alpha} \sum_{z=0}^{N} \int dF^*(\lambda)\, P(z\,|\lambda) \Big[\lambda \ln \frac{\lambda}{b(z)} + (1 - \lambda) \ln \frac{(1 - \lambda)}{1 - b(z)}\Big]\Big|_{b(z) = b_{F^* + \alpha G}(z)}$$

$$= \sum_{z=0}^{N} - \int dF^*(\lambda)\, P(z\,|\lambda) \Big[\frac{\lambda}{b_{F^*}(z)} - \frac{1 - \lambda}{1 - b_{F^*}(z)}\Big] \lim_{\alpha \to 0} \frac{\partial b_{F^* + \alpha G}(z)}{\partial \alpha}$$

$$= \sum_{z=0}^{N} - \lim_{\alpha \to 0} \frac{\partial b_{F^* + \alpha G}(z)}{\partial \alpha} \frac{\int dF^*(\lambda) P(z|\lambda)(\lambda - b_{F^*}(z))}{b_{F^*}(z)(1 - b_{F^*}(z))}$$

$$= 0 \qquad (2.18)$$

since the integral is zero by the definition of $b_{F^*}(z)$.

Substituting (2.17), (2.18) in (2.16), we get

$$\delta\bar{\psi} = \int dG(\lambda)\left[r(\lambda, b_{F^*}) + \mu + c(\lambda)\right] \qquad (2.19)$$

Since $dG(\lambda)$ is arbitrary, $\delta\bar{\psi} \leq 0$ for all $dG(.)$ if and only if

$$r(\lambda, b_{F^*}) + \mu + c(\lambda) = 0 \qquad (2.20)$$

$$\begin{aligned}
R(F^*, b^*) &= \int dF^* r(\lambda, b_{F^*}) \\
&= \int dF^*(-\mu - c(\lambda)) \\
&= -\mu \text{ by } (2.13) \\
R(F, b^*) &= \int dF(\lambda) r(\lambda, b_{F^*}) \\
&= \int dF(\lambda)(-\mu - c(\lambda)) \\
&\leq -\mu \text{ , since } c(\lambda) \geq 0 \\
&\leq R(F^*, b^*), \text{ Q.E.D.}
\end{aligned}$$

III. THE STRUCTURE OF THE OPTIMAL PRIOR C.D.F. F*

In Theorem 2, we will consider the structure of the "worst" C.D.F. F*.

Theorem 2: Consider the game $(\mathcal{F}, \mathcal{B}, R)$. The "worst" C.D.F. F* is a stair case function with, at most, N steps, N being the size of the observation sample.

Proof of Theorem 2:

$$\text{Let } \rho_i(F) = \int \lambda^i dF(\lambda) \qquad (3.1)$$

By definition of $b_F(.)$

$$\begin{aligned}
b_F(z) &= \int \lambda \lambda^z (1 - \lambda)^{N-z} dF(\lambda) / \int \lambda^z (1 - \lambda)^{N-z} dF(\lambda) \\
&= f_z(\rho_1(F), \rho_2(F), \ldots, \rho_{N+1}(F)), \quad z = 0, 1, \ldots, N
\end{aligned} \qquad (3.2)$$

where $f_z(.)$ is some function of the indicated arguments.
Similarly,

$$\int \lambda P(z|\lambda) dF(\lambda) = f_{N+1}(\rho_1(F), \ldots, \rho_{N+1}(F)) \qquad (3.3)$$

$$\begin{aligned}
\psi(F) &\overset{\Delta}{=} R(F, b_F) \\
&= \sum_{z=0}^{N} \int dF(\lambda) P(z|\lambda) \left[\lambda \ln \frac{\lambda}{b_F(z)} + (1 - \lambda) \ln \frac{1 - \lambda}{1 - b_F(z)}\right]
\end{aligned}$$

$$= \int dF(\lambda)[\lambda \ln\lambda + (1 - \lambda)\ln(1 - \lambda)]$$

$$- \sum_{z=0}^{N} \int dF(\lambda)P(z|\lambda)[\lambda \ln b_F(z) + (1 - \lambda)\ln(1 - b_F(z)] \tag{3.4}$$

$$= -\int dF(\lambda)H(\lambda) + g(\rho_1(F),\ldots,\rho_{N+1}(F)) \tag{3.5}$$

where $H(\lambda) = -\lambda\ln\lambda - (1 - \lambda)\ln(1 - \lambda)$ (3.6)

and g represents the second term in (3.4) after simplification using the relations (3.2) and (3.3).

We will show in [3] that $\psi(F)$ is a concave functional of F. Since $\rho_i(\alpha F_1 + (1 - \alpha)F_2) = \alpha\rho_i(F_1) + (1 - \alpha)\rho_i(F_2)$ by (3.1), the concavity of $\psi(F)$ in F implies the concavity of the function $g(\rho_1,\ldots,\rho_{N+1})$ in the variables $\rho_1,\ldots,\rho_{N+1}$. Let F* be the supremizing function of $\psi(F)$ subject to the constraints in (3.7) and (3.8) and let $\rho^*_1,\ldots,\rho^*_{N+1}$ be the corresponding moments, $\rho^*_i = \rho_i(F^*)$.

$$\int dF(\lambda) = 1 \tag{3.7}$$

$$dF(\lambda) \geq 0 \;\forall\; \lambda \in [0,1] \tag{3.8}$$

Define an augmented function $\bar{\psi}$, involving the functional ψ, a set of parameters $\rho_1,\ldots,\rho_{N+1}$ (not necessarily the moments) and the multipliers $\mu_0,\ldots,\mu_{N+1}$.

$$\bar{\psi}(F,\rho_1,\ldots,\rho_{N+1}) = -\int dF(\lambda)H(\lambda) + g(\rho_1,\ldots,\rho_{N+1})$$

$$+ \sum_{i=1}^{N+1} \mu_i\left(\int\lambda^i\, dF(\lambda) - \rho_i\right) \tag{3.9}$$

Note that $\bar{\psi}(F)$ is linear in F and concave in the variables $\rho_1,\ldots,\rho_{N+1}$. By the theory of Lagrange multipliers, there exists a set of multipliers $\mu_0,\ldots,\mu_{N+1}$ so that (3.10) is true for all F obeying only (3.7) and (3.8) and all $\rho_1,\ldots,\rho_{N+1}$.

$$\bar{\psi}(F^*,\rho^*_1,\ldots,\rho^*_{N+1}) \geq \bar{\psi}(F,\rho_1,\ldots,\rho_{N+1}) \tag{3.10}$$

Rearranging $\bar{\psi}$ in (3.9) we get

$$\bar{\psi} = \int dF(\lambda)\Big[\underbrace{-H(\lambda) + \sum_{i=1}^{N+1}\mu_i\lambda^i}_{= h(\lambda)}\Big] - \sum_{i=1}^{N+1}\mu_i\rho_i + g(\rho_1,\ldots,\rho_{N+1}) \tag{3.11}$$

Let $h_o = \sup_{0 \leq \lambda \leq 1}.\; h(\lambda)$

To determine the points λ at which $h(\lambda) = h_o$, note that

$$\frac{dh}{d\lambda} = \sum^{N}(j + 1)\mu_{j+1}\lambda^j - \ln(\lambda/1 - \lambda)$$

Since $dh/d\lambda$ is a sum of a N^{th} degree polynomial in λ and a function $\ln(\lambda/1 - \lambda)$ which is monotonically increasing, the function $dh(\lambda)/d\lambda$ can have at most N zeroes in the interval $0 < \lambda < 1$. Consequently, the function $h(\lambda)$ can assume the value h_o at utmost, N + 2 points, the N interior points mentioned above and the 2 boundary points $\lambda = 0$, and $\lambda = 1$. Let these points be labelled $\lambda_0, \ldots, \lambda_{N+1}$.

Now consider the expression (3.11) for $\bar{\psi}$ in conjunction with (3.10). Since $dF(\lambda)$ occurs linearly in (3.11), F* could be a maximizing function as defined in (3.10) among F obeying (3.7), only if

$$\begin{aligned} dF^*(\lambda) &\geq 0 \quad \forall\, \lambda = \lambda_0, \ldots, \lambda_{N+1} \\ &= 0 \quad \text{elsewhere} \qquad \text{Q.E.D.} \end{aligned} \tag{3.12}$$

Check on the result:

In the above proof, we used the theory of Lagrange multipliers to invoke the existence of the multipliers obeying (3.10). In this section we will display these multipliers explicitly for the case N = 1. Of course, the similar procedure can be used for higher N.

The candidate for F* when N = 1 is displayed here

$$F^*(\lambda) = 0.3\, u(\lambda) + 0.4\, u(\lambda - 0.5) + 0.3\, u(\lambda - 1)$$

where u(.) is the unit step function, $u(\lambda) = 0$ for $\lambda < 0$ and 1 elsewhere. The corresponding moments are

$$\rho^*_1 = \int \lambda dF^*(\lambda) = 0.5, \; \rho^*_2 = \int \lambda^2 dF^*(\lambda) = 0.4.$$

The corresponding candidates μ_i are: $\mu_1 = 4\ln 2$, $\mu_2 = -4\ln 2$.
We need to show that

$$\bar{\psi}(F^*, \rho^*_1, \rho^*_2) \geq \bar{\psi}(F, \rho_1, \rho_2)$$

for all F obeying (3.7) and (3.8) and all ρ_1, ρ_2. We will first evaluate the function g in ψ from (3.4) and (3.5).

$$\begin{aligned} -g(\rho_1, \rho_2) &\triangleq -\int dF(\lambda)(1 - \lambda)\left[\lambda \ln\left(1 - \frac{\rho_2}{\rho_1}\right) + (1 - \lambda) \ln \frac{\rho_2}{\rho_1}\right] \\ &\quad - \int dF(\lambda)\lambda\left[\lambda \ln \rho_2/\rho_1 + (1 - \lambda) \ln(1 - \rho_2/\rho_1)\right] \\ &= 2(\rho_2 - \rho_1) \ln\left(1 - \frac{\rho_2}{\rho_1}\right) + (2\rho_1 - 2\rho_2 - 1) \ln(\rho_2/\rho_1) \end{aligned}$$

$$H(\lambda) = [-\lambda \ln\lambda + (1 - \lambda) \ln(1 - \lambda)]$$

By (3.11)

$$\bar{\psi}(F) = \int dF(\lambda)[-H(\lambda) + \mu_1\lambda + \mu_2\lambda^2] + g(\rho_1, \rho_2) - \mu_1\rho_1 - \mu_2\rho_2 \tag{3.13}$$

$$\text{Let } g'(\rho_1, \rho_2) = g(\rho_1, \rho_2) - \mu_1\rho_1 - \mu_2\rho_2 \tag{3.14}$$

$$\frac{\partial g'(\rho_1,\rho_2)}{\partial \rho_1} = -2 - 2\ln(\rho_1 - \rho_2) + 1/\rho_1 + 2\ln\rho_2 - \mu_1$$

$$= 0, \text{ at } \rho_1 = \rho^*_1,\ \rho_2 = \rho^*_2 \qquad (3.15)$$

Similarly,

$$\frac{\partial g'(\rho_1,\rho_2)}{\partial \rho_2} = 0 \quad \text{at } \rho_1 = \rho^*_1,\ \rho_2 = \rho^*_2 \qquad (3.16)$$

One can also check the matrix of second derivatives is positive definite at $\rho_1 = \rho^*_1$ and $\rho_2 = \rho^*_2$.

Let $h(\lambda) = -H(\lambda) + \mu_1\lambda + \mu_2\lambda^2$

= The coefficient of $dF(\lambda)$ in (3.13).

$h(\lambda)$ assumes the maximum value of 0 at three points only in $0 \leq \lambda \leq 1$ namely $\lambda_0 = 0$, $\lambda_1 = 0.5$, $\lambda_2 = 1.0$. In view of the above facts, (3.15) and (3.16),

$$\bar{\psi}(F^*,\rho^*_1,\rho^*_2) \geq \bar{\psi}(F,\rho_1,\rho_2)$$

for any ρ_1,ρ_2 and F obeying (3.7) and (3.8).

Discussion:

The highlight of this section is the stair cased form of the optimal C.D.F. F*. This is unusual considering the facts that the loss function was a differentiable function (except at the boundary) and the solutions of most of the statistical minimax problems [1] result in smooth prior C.D.F.'s. One of the plausible reasons for this feature may be that the loss function used here L(,) is asymmetric in the 2 variables.

The second point that has to be emphasized is that the optimal C.D.F. F* depends on N, since N is a parameter in the game $(\mathcal{S},\mathcal{B},R)$. We have tabulated the F* function with N = 0,1 and 2.

$N = 0$: $F^*(\lambda) = 0.5\,u(\lambda) + 0.5\,u(\lambda - 1)$

$N = 1$: $F^*(\lambda) = 0.3\,u(\lambda) + 0.4\,u(\lambda - 0.5) + 0.3\,u(\lambda - 1)$

$N = 2$: $F^*(\lambda) = .266\,u(\lambda) + 0.468\,u(\lambda - 0.5) + .266\,u(\lambda - 1)$

IV. APPROXIMATION TO F* AND b* FOR LARGE N

When N is small, we can directly find F* by Theorem 2. When N is even as large as 8 or 10, the computation of F* directly becomes cumbersome. Hence we have to develop indirect methods of computation.

In approximating the C.D.F. F* we need some criterion function. We note that one of the important properties of the estimate b* derived from F* is the following:

$$r(\lambda,b_{F^*}) \leq \psi(F^*) \quad \forall\ \lambda$$

We know [3] that for large N

$$\psi(F^*) = O(1/N)$$

Hence we look for an approximation $\bar{F}$ to F* such that the following inequality is true:

$$|r(\lambda,b_{\bar{F}}) - \psi(F^*)| = O(1/N)$$

i.e., for large N, the difference in the risk involved between the estimates b_{F*} and $b_{\overline{F}}$ is negligible. Such an approximation $\overline{F}$ is given below [3].

$$\overline{F}(\lambda) = \frac{1}{N+2} \sum_{i=0}^{N+1} u(\lambda - \lambda_i), \quad \lambda_i = \frac{i}{N+1}$$

i.e., $\overline{F}$ is a uniformly stepped stair case function.

It is easy to see that with large N, the properties of the estimate b_F will be essentially the same as the Bayes estimate of λ with a uniform prior CDF. However, the properties of these two estimates will be drastically different for small N.

V. INFORMATION THEORETIC INTERPRETATION OF F*

Consider a sequence of (N + 1) independent coin tossings, the probability of occurrence of a head in a trial being λ. Let the number of heads in N trials be denoted by Z and label the results of the (N + 1) trial as $Y \in \{1,0\}$, 1 denoting head. How can we measure the uncertainty about the occurrence of Y caused by the ignorance of λ, given the history Z?

The total uncertainty about Y is made of two parts, namely (i) the uncertainty about Y caused by the random nature of Y, even when λ is known and (ii) the uncertainty about Y caused by the ignorance of λ alone. We are interested in measuring the uncertainty in part (ii) alone. Note the knowledge of the history Z helps in reduction of the uncertainty in part (ii) only. The observation Z cannot influence the uncertainty in part (i) since Y and Z are statistically independent if λ is unknown.

Since λ is unknown, we can regard it as a random variable with prior C.D.F. F. If we imbed the problem in a repetative framework, then the prior C.D.F. F is nothing but the corresponding cumulative frequency distribution. We will have more to say about the choice of F.

We will measure all the uncertainties by the corresponding entropies. We will first find the average uncertainty in part (i):

The uncertainty about Y caused by its randomness given λ = $H(\lambda)$

$$= -[\lambda \ln\lambda + (1 - \lambda) \ln(1 - \lambda)]$$

Averaging this expression over λ we get

J_1 = The average uncertainty about Y caused by its randomness

$$= \int H(\lambda) dF(\lambda)$$

Now consider the total average uncertainty about Y caused by both the randomness of Y and the ignorance of λ. Let $\hat{\lambda}_F(Z)$ be the Bayes estimate of λ based on the prior C.D.F. F, i.e., given the observation Z, $\hat{\lambda}_F(Z)$ represents the posterior probability of the variable Y assuming the value 1 (or head). Hence, entropy of this posterior probability distribution

$$= H(\hat{\lambda}_F(Z))$$

$$= -[\hat{\lambda}_F(Z) \ln\hat{\lambda}_F(Z) + (1 - \hat{\lambda}_F(Z) \ln(1 - \hat{\lambda}_F(Z))]$$

Averaging this entropy over both the random variables Z and λ, we get the total average uncertainty J_2

$$J_2 = \int dF(\lambda) \sum_{z=0}^{N} P(z|\lambda) H(\hat{\lambda}_F(z))$$

$J \overset{\Delta}{=}$ The average uncertainty about Y caused by the ignorance of λ alone, given Z

$$= \text{The total uncertainty caused by unknown } \lambda \text{ and the randomness of } Y - \text{Average uncertainty caused by the randomness of } Y \text{ only}$$

$$= J_2 - J_1$$

$$= \int dF(\lambda) \sum_{z=0}^{N} P(z|\lambda)[H(\hat{\lambda}_F(z)) - H(\lambda)]$$

$$= \int dF(\lambda) \sum_{z=0}^{N} P(z|\lambda)[-\hat{\lambda}_F(z)\ln\hat{\lambda}_F(z) + \lambda\ln\lambda - (1 - \hat{\lambda}_F(z))\ln(1 - \hat{\lambda}_F(z))$$

$$+ (1 - \lambda)\ln(1 - \lambda)]$$

$$= \int dF(\lambda) \sum_{z=0}^{N} P(z|\lambda)\left[\lambda\ln\frac{\lambda}{\hat{\lambda}_F(z)} + (1 - \lambda)\ln\frac{1 - \lambda}{1 - \hat{\lambda}_F(z)}\right]$$

$$\overset{\Delta}{=} \psi(F)$$

Thus the functional $\psi(F)$ encountered earlier can be interpreted as an uncertainty function. Note the functional $J(F)$ is also termed as the Shannon mutual information between the output variable Y and the input λ conditioned by the observation Z.

If the C.D.F. F is unknown, a "safe" strategy is to choose that function which maximizes the uncertainty functional. i.e., F* can be interpreted to be a least informative prior C.D.F.

We have already referred to the fact that $\psi(F^*) = \mathcal{O}(1/N)$ for large N. As the history of the observations accumulate, the marginal decrease in the uncertainty caused by the latest observation is only of the order $\mathcal{O}(1/N^2)$. As the duration N of the observations Z tend to ∞, the uncertainty caused only by the ignorance of λ vanishes since effectively we can recover the unknown λ from the history of the observations.

VI. DISCUSSION

We will discuss the properties of the estimate $\bar{b}$. An important aspect of the estimate $\bar{b}$ is that it is derived from a prior C.D.F. $\bar{F}$ which depends on the sample size N. This is not unexpected since the C.D.F. is derived from a game which is indexed by the integer. Still, the prior C.D.F.'s occuring in the Bayesian literature are all usually independent of N.

We will compare the properties of the estimate $\bar{b}$ with two other estimates, namely the maximum likelihood estimate $b_M(z) = z/N$ and the minimax estimate derived from quadratic loss function, $b_Q = (z + \sqrt{N}/2)/(z + N)$.

The estimates $\bar{b}$ and b_M are asymptotically efficient in the sense of Rao [6], whereas the minimax estimate b_Q is not asymptotically efficient [2].

The small sample behavior of the three estimates can be compared using two risk functions, the quadratic risk function $r_1(\lambda,b)$ in (6.1) and the risk function $r(\lambda,b)$ defined in (2.3)

$$r_1(\lambda,b) = \sum_{z=0}^{N} P(z|\lambda)(\lambda - b(z))^2 \tag{6.1}$$

The graph of $r_1(\lambda,b)$ versus of λ is plotted in Figure 2 for N = 10 and N = 20. Note that the $r_1(\lambda,b)$ characteristic for the ML estimate b_M is complimentary to that of $\bar{b}$. Moreover, if we compare the three estimates by using the mean risk $\bar{r}_1 = \int r_1(\lambda,b)d\lambda$, then the estimate $\bar{b}$ has the smallest mean risk in comparison with both the other two estimates b_M and b_Q. A similar property is true for $\bar{b}$, with the risk function $r(\lambda,\bar{b})$. Consequently, if we are asked to choose one estimate for both large and small sample estimation, the estimate $\bar{b}$ appears to be the answer.

All the remarks made here are valid only for the binomial distribution. However, the relevance of these results for multinomial distribution is discussed in the author's report [3] (TR EE 72-31, Purdue University).

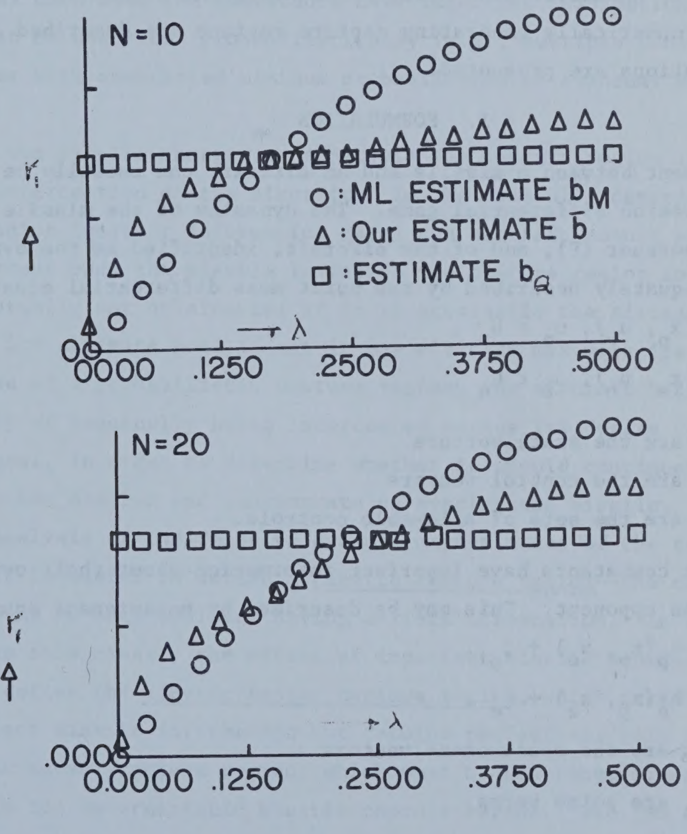


Fig. 2: The graph of $r_1(\lambda,b)$ vs. λ for $b = b_M, b_Q, \bar{b}$

THE GENERATION AND APPLICATION OF CAPTURE REGIONS IN AIRCRAFT/MISSILE ENGAGEMENTS

DAVID M. SALMON
LEWIS MEIER III
Systems Control, Inc.
260 Sheridan Ave.
Palo Alto, Ca. 94306

ABSTRACT

This paper introduces the idea of a capture region in the context of analyzing aircraft/missile engagements. The capture region is the set of initial states of the vehicles from which the missile can eventually intercept the aircraft no matter what evasive maneuvers the aircraft makes. Techniques for numerically generating capture regions are described and some example applications are presented.

1. FORMULATION

An engagement between a missile and an aircraft can usefully be regarded as a pursuit-evasion differential game. The dynamics of the missile, identified as the pursuer (P), and of the aircraft, identified as the evader (E), are usually adequately described by the point mass differential equations of form

$$\begin{aligned} \dot{x}_p &= f_p(x_p, u_p),\ u_p \in U_p\ , \\ \dot{x}_e &= f_e(x_e, u_e),\ u_e \in U_e\ , \end{aligned} \tag{1}$$

where x_p, x_e are the state vectors
u_p, u_e are the control vectors
U_p, U_e are the sets of allowable controls.

In general, the combatants have imperfect information about their own states and those of the opponent. This may be described by measurement equations of form

$$\begin{aligned} z_p &= h_p(x_p, x_e) + \xi_p\ , \\ z_e &= h_e(x_p, x_e) + \xi_e\ , \end{aligned} \tag{2}$$

where z_p, z_e, are the measurement vectors
ξ_p, ξ_e, are noise terms.

The ultimate goals of the aircraft and missile might be respectively to attack and defend a particular site (which may itself be moving), or to penetrate and deny penetration of a defense barrier. A subsidiary goal is respectively for the aircraft to avoid being intercepted (killed) by the missile and for the missile to achieve an intercept of the aircraft. Intercept is assumed to occur if the missile achieves adequate proximity to the aircraft, i.e., if $\emptyset_1(x_p, x_e) = 0.$ (3)

2. CAPTURE REGION DEFINITIONS

Because a missile normally has less endurance than an aircraft, it is obvious that some initial states of the vehicles exist from which a properly controlled aircraft can evade the missile. It is also evident that usually there are some initial states (e.g., some of the states almost satisfying Eq. (3)) from which a properly controlled missile can eventually intercept the aircraft no matter what maneuvers the aircraft makes. Accordingly the capture region is defined as the set of initial states of the combatants from which the missile, if properly controlled, can eventually achieve the intercept condition of Eq. (3) no matter what maneuvers the aircraft makes. In the general case when the combatants have imperfect information, the capture region must be described probabilistically (e.g., multiple contours of relative states with associated minimum probabilities of eventual aircraft interception).

Once the missile manages to penetrate the capture region its goal is to expedite interception of the aircraft. In the case of a deterministic capture region (perfect information case) the aircraft cannot escape an eventual intercept once the missile is inside the capture region so its ultimate goal is normally not attainable; if it is attainable the aircraft may elect to pursue its ultimate goal, if not it may elect to maximize its survival time. In the case of a probabilistic capture region, the aircraft must balance its probability of eventually being intercepted versus its desire to achieve its ultimate goal, in order to determine whether it should continue its mission or abandon the mission and concentrate on evading the missile.

For analysis of the relative dynamic capabilities of the two combating vehicles it is useful to define a kinetic capture region--the capture region corresponding to both vehicles having perfect information. In order, for example, to show clearly the effect of imperfect missile sensing it is appropriate to define the kinetic-sensor capture region, which takes into account the imperfect missile information but retains perfect aircraft information. The kinetic-sensor capture region, which must be described probabilistically, lies within the deterministic kinetic capture region. The two types of capture region should be generated and compared when performing design tradeoffs between the dynamic and sensing capabilities of the missile. In some cases it may be useful to go even further and define, for example, a kinetic-sensor-computation capture region that shows the capture region actually obtained with a particular set of real time estimation and control algorithms as limited by the available real time data processing capability.

In practice, the kinetic capture region is much easier to obtain numerically than the various probabilistic capture regions. However, the latter can usually be obtained with adequate accuracy by making selected runs of a computer program simulating the engagement. The following section describes numerical procedures for generating kinetic capture regions.

3. NUMERICAL TECHNIQUES

Conceptually a kinetic capture region is determined by following two steps (see Reference 1 for a more detailed discussion of this topic):

1. The sets of positions that the missile and aircraft can reach in a sequence of given times from given initial conditions are computed using minimum-time trajectories. An example set of such reachable positions for a particular aircraft is shown in Fig. 1.
2. These two sets of reachable positions are combined to determine the relative states from which the aircraft cannot evade an eventual intercept by the missile.

In practice it is usually best to not generate the entire sets of reachable positions (Step 1) before combining them (Step 2). Consider, for example, an engagement between an aircraft and an aerodynamically controlled missile, with reachable positions having a similar form to those of Fig. 1. For simplicity, assume both vehicles maneuver in a plane and assume the aircraft is intercepted if the distance between the vehicles is ℓ and the missile has an adequate speed advantage,* i.e., Eq. (3) is

$$(x_{p_1} - x_{e_1})^2 + (x_{p_2} - x_{e_2})^2 + (x_{p_3} - x_{e_3})^2 = \ell^2$$

$$v_p \geq kv_e \text{ , } k > 1.0 \text{ ,} \tag{4}$$

where x_{p_i} , x_{e_i} are the cartesian vehicle positions

v_p , v_e are the magnitudes of the vehicle velocities.

Given a missile with maneuver capabilities similar to those of Fig. 1 and given the intercept conditions (Eq. (4)) there are two ways the aircraft may conceivably evade the missile: by maneuvering inside the maximum-turn envelope of the missile reachable positions (assuming the missile does not have the endurance to make a second pass which threatens the aircraft); or by outrunning the missile. The former is called "cut-inside escape" and the latter "tail-chase escape." To find points on the boundary of the capture region it is necessary to find initial aircraft/missile states from which the aircraft cannot quite achieve either a cut-inside or a tail-chase escape.

---

*The requirement for a missile speed advantage may result from a detailed end-game analysis showing that the missile can follow last-minute evasion maneuvers of the aircraft only if it has a sufficient speed advantage over the aircraft.

Consider first the problem of finding points on the capture region boundary corresponding to cut-inside escape. For such points the missile equations are integrated forward from a specified initial position for a specified period of time with the control corresponding to a hard turn to reach a point such as C in Fig. 2, for example. Now it is possible to determine the conditions at C such that the aircraft does not quite achieve a cut-inside escape. First, the aircraft/missile separation must be exactly equal to ℓ. Second, if the aircraft is travelling straight before reaching C, its velocity must be perpendicular to the missile/aircraft relative velocity, so that the aircraft/missile separation is exactly ℓ when the missile is at C but greater than ℓ both before and after that time. A modified verge-of-capture condition applies if the aircraft is turning before it reaches C. (These conditions correspond to the transversality conditions of the differential game.) Starting from this verge-of-capture condition the aircraft equations are integrated backward in time for the length of time it took the missile to reach C. Points such as E_1, E_2, E_3 and E_4 in Fig. 2 are obtained by varying the length of turn of the aircraft during the backward integration. For the given initial missile position, points E_1 through E_4 represent initial aircraft states from which the aircraft cannot quite escape by cutting inside the envelope of missile reachable positions, i.e., they are points on the cut-inside boundary. Points on the chase-escape boundary are found in a similar manner with two differences: first, the missile equations are integrated forward with an initial turn, followed by non-turning flight, until the missile speed has fallen to k times the aircraft speed, and second, the aircraft final heading is in the same direction as that of the missile final heading.

By combining the appropriate points generated in the above manner, cross-sections of the capture region can be generated for particular initial aircraft headings as shown in Figs. 3 and 4. The internal structure of the capture region can be found by varying the parameter k which specifies the speed advantage of the missile at intercept. Of particular interest in Figs. 3 and 4 are the curves that divide regions of different turn direction for one or both vehicles. For example, in Fig. 3 if the aircraft is above the line labelled "evader and pursuer dispersal surface" the aircraft turns right while the missile turns left. If the aircraft is below the line it turns left while the missile turns right.

4. EXAMPLE APPLICATIONS

Many questions arisen in the design and analysis of an aircraft or missile engagement. For example, in the case of a missile system intended to protect a particular site, three fundamental questions are:

1. For given missile and aircraft capabilities what are the positions of the site relative to the missile system such that the site can be protected from the aircraft?

2. For a missile system with particular overall capabilities versus a specified aircraft threat, what is an appropriate (i.e., cost effective) balance of dynamic, sensing, and weapon (warhead) capabilities?

3. For a missile system or an aircraft system with given dynamic and sensing capabilities what policies should be employed in real time to control the vehicle and its weapon?

The two brief examples which follow illustrate how capture region analysis can be applied to answer such questions.

Analysis of a Surface to Air Missile System

Consider the problem of designing a surface-to-air missile system intended to protect sites from the attack of hostile aircraft.

Two key questions are:

1. When must the system launch a missile against a given aircraft in order to protect a given site from that aircraft?

2. Given that a missile is launched, is it possible for the aircraft to evade interception by the missile (perhaps at the cost of failing to achieve its mission)?

Figure 5 shows example launch boundaries--a missile must be launched before an aircraft arrives at each such boundary in order to protect each of the sites marked 1, 2, 3, 4 and the missile site itself. The figure also shows the capture region--a set of aircraft positions at missile launch from which the aircraft cannot evade eventual interception. The figure has been generated for specific aircraft and missile capabilities on the assumption that the aircraft will maneuver in the horizontal plane (a reasonable assumption for a low-altitude engagement). Two modes of aircraft escape from interception were found to be dominant when generating the capture region--chase-escape (as discussed in the previous section) and maneuvers carrying the aircraft outside the surveillance limit, shown in Fig. 5, before the intercept.

The figure shows, for example, that while Site 1 can be protected, it will normally be necessary to launch a missile against an aircraft threatening Site 1 before the aircraft is inside the capture region, so there exists an evasion maneuver that the aircraft may employ after missile launch which enables it to escape intercept. Site 4, however, can be protected even if the missile is not launched until the aircraft is within the capture region. Thus each missile launched to protect Site 4 should successfully intercept an aircraft. Information of this type is valuable for performing design tradeoff studies, deployment studies, and for incorporating into the real-time criteria to be used for launching the missile.

Similar analyses can, of course, be made to identify the best tactics and real time decisions for the aircraft. For this purpose it is usually more convenient to draw the capture region relative to the aircraft position rather than relative to the initial missile position as in Fig. 5.

Guidance Law Design

A second example application of capture regions is in the design of real time control laws. This topic is treated in greater detail in Reference 2. Consider the aircraft/missile system yielding the capture region cross sections of Figs. 4 and 5. The internal structure of the capture regions shows the best aircraft strategies for various relative aircraft/missile states. In particular, the evader dispersal surface separates the regions where the aircraft should evade to the right or to the left. A real time guidance law for the missile can be devised based on first determining the aircraft position relative to the dispersal surface, then determining the best evasion trajectory for the aircraft, and finally guiding the missile towards an intercept of this trajectory. If the aircraft, in fact, follows a non-optimal trajectory the interception task is made easier. The feedback guidance law by continually re-determining the best evasion trajectory for the aircraft gradually modifies the missile trajectory in response to whatever trajectory the aircraft follows. Figure 6 compares missile trajectories resulting from two guidance laws attempting to intercept an aircraft which evades correctly and starts from a position just inside the capture region. The missile when guided by a proportional navigation guidance law does not achieve an intercept by the time its speed advantage is lost, but the same missile when guided by the new guidance law just achieves an intercept.

5. CONCLUSIONS

The capture region provides a convenient way of representing the capability of a missile vis-a-vis an aircraft. It is valuable in relating subsystem performance levels to overall system effectiveness. In this regard the capture region can be used to structure an overall missile/aircraft system analysis problem into a convenient hierarchy of three levels of analysis

1. Weapon System Analysis--At this level the effectiveness of the missile warhead versus the aircraft is analyzed in detail in order to determine the aircraft/missile states that define an aircraft intercept. (The effectiveness of the aircraft weapon versus any site it may attack can be determined similarly.)
2. Capture Region Analysis--At this level it is necessary to determine the capture region--the set of initial states from which the missile can eventually achieve the intercept conditions identified by the weapon system analysis.
3. Overall System Analysis--At this level it is necessary to determine the best strategies for the combatants to follow and their effectiveness. For example, what aircraft trajectories (if any) allow the aircraft to attack its target while not allowing the missile to penetrate the capture region?

When both vehicles have perfect information it is relatively straightforward to develop a computer program capable of numerically generating kinetic capture regions. If one or both vehicles has imperfect information a simulation approach can be used to identify the perturbations to the kinetic capture region caused by the imperfect information.

Example applications have been given which indicate how kinetic capture regions can be used both during the initial design of an aircraft/missile system to determine basic characteristics of the vehicles, as well as later in the design to determine real-time algorithms to guide the vehicles or to allocate resources.

6. REFERENCES

1. David M. Salmon, "Policies and Controller Design for a Pursuing Vehicle," IEEE Trans. on Automatic Control, Vol. AC-14, No. 5, October 1969.

2. Lewis Meier, III, "A New Technique for Solving Pursuit-Evasion Differential Games," IEEE Trans. on Automatic Control, Vol. AC-14, August 1969.

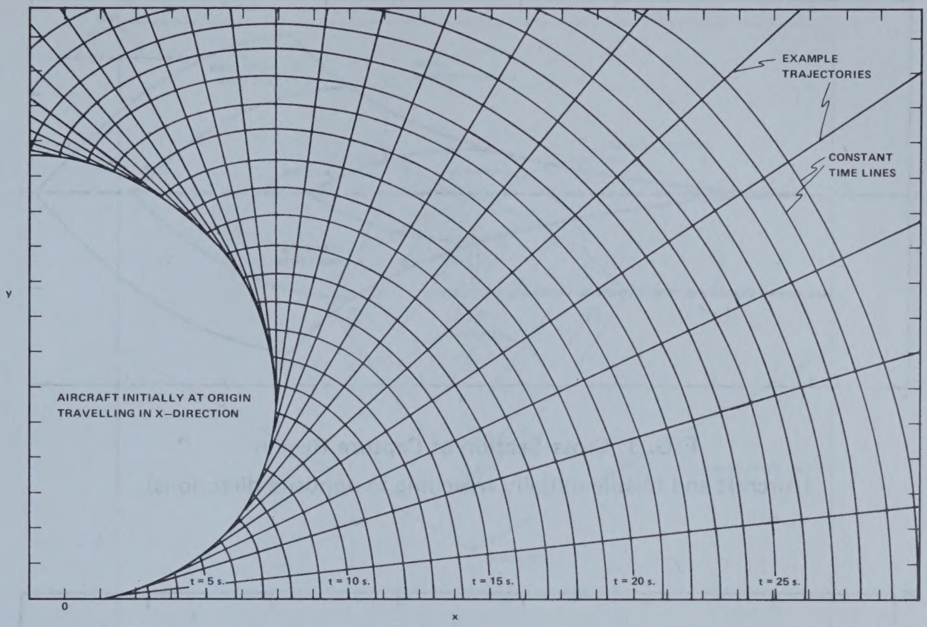


FIG. 1: Example Set of Aircraft Reachable Positions

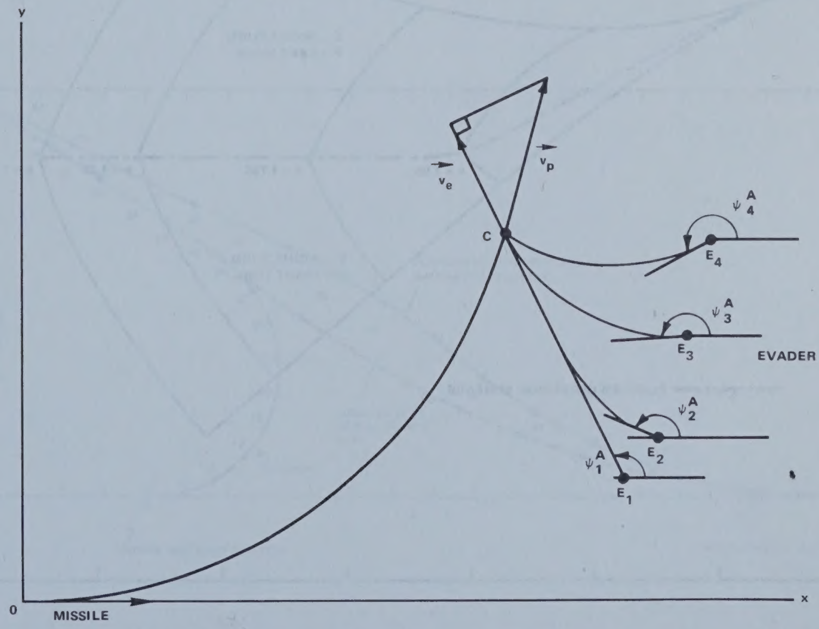

y
$\vec{v}_e$
$\vec{v}_p$
C
ψ_4^A
E_4
ψ_3^A
E_3
EVADER
ψ_2^A
E_2
ψ_1^A
E_1
0
MISSILE
x

FIG. 2: Determination of the Cut–Inside Boundary

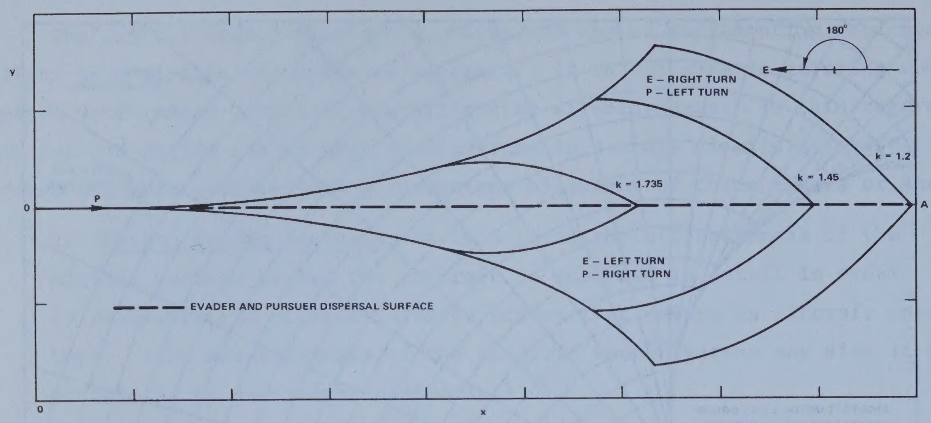

180°
E
y
E – RIGHT TURN
P – LEFT TURN
k = 1.2
k = 1.45
k = 1.735
P
0
A
E – LEFT TURN
P – RIGHT TURN
EVADER AND PURSUER DISPERSAL SURFACE
0
x

FIG. 3: Cross-Section of Capture Region
(Aircraft and Missile initially travelling in opposite directions).

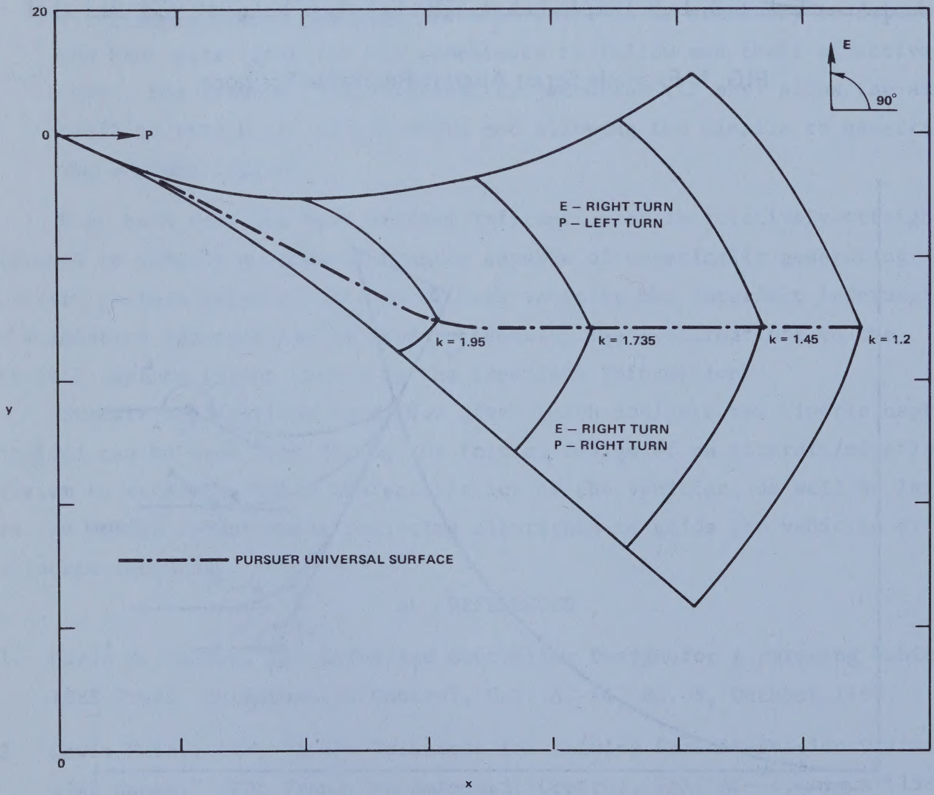

20
E
90°
0
P
E – RIGHT TURN
P – LEFT TURN
k = 1.95
k = 1.735
k = 1.45
k = 1.2
y
E – RIGHT TURN
P – RIGHT TURN
PURSUER UNIVERSAL SURFACE
0
x

FIG. 4: Cross-Section of Capture Region
(Aircraft and Missile initial headings differ by 90°).

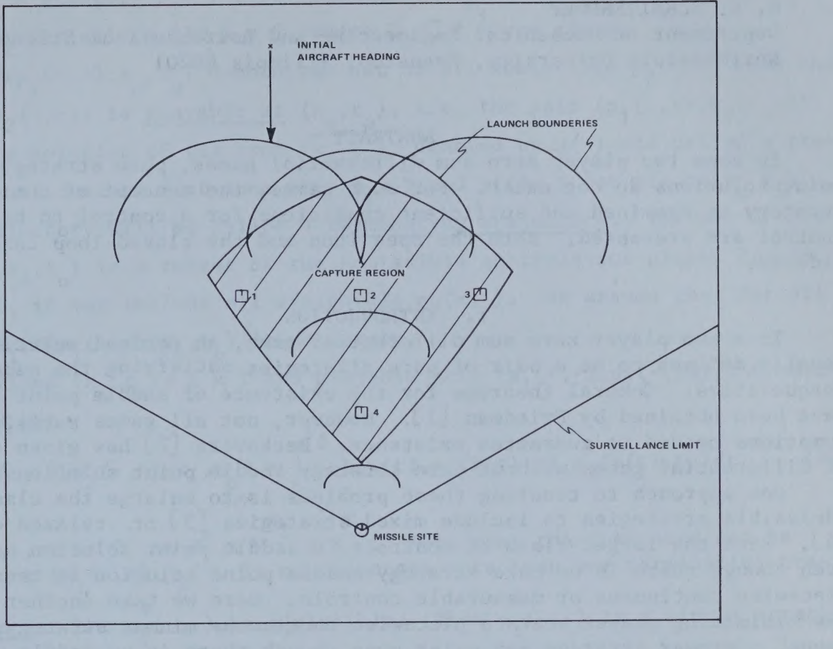

INITIAL AIRCRAFT HEADING
LAUNCH BOUNDERIES
CAPTURE REGION
1
2
3
4
SURVEILLANCE LIMIT
MISSILE SITE

FIG. 5: Capture Region and Launch Boundaries for a Surface to Air Missile System

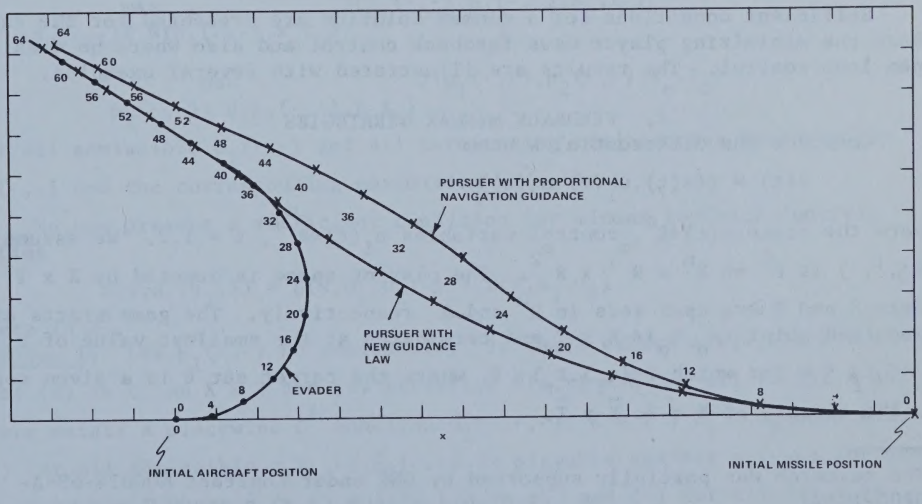


FIG. 6: Missile Response with Two Guidance Laws to Attempted Tail-Chase Escape (Numbers on trajectories give time).

DIFFERENTIAL GAMES WITHOUT PURE STRATEGY SADDLE POINT SOLUTIONS[1]

W. E. SCHMITENDORF
Department of Mechanical Engineering and Astronautical Sciences,
Northwestern University, Evanston, Illinois 60201

ABSTRACT

In some two player zero sum differential games, pure strategy saddle point solutions do not exist. For such games, the concept of a minmax strategy is examined and sufficient conditions for a control to be a minmax control are presented. Both the open loop and the closed loop case are considered.

1. INTRODUCTION

In a two player zero sum differential game, an optimal solution is usually defined to be a pair of pure strategies satisfying the saddle point inequalities. General theorems for the existence of saddle point solutions have been obtained by Friedman [1]. However, not all games satisfy the assumptions needed to guarantee existence. Berkovitz [2] has given examples of differential games without pure strategy saddle point solutions.

One approach to treating these problems is to enlarge the class of admissible strategies to include mixed strategies [3] or relaxed controls [4]. With the larger class of controls, a saddle point solution may exist even though there is no pure strategy saddle point solution in the class of piecewise continuous or measurable controls. Here we take another approach. The minimizing player seeks a piecewise continuous minmax strategy. In some games, a minmax strategy may exist even though there is no saddle point solution. With a minmax strategy, the minimizing player can guarantee the minmax value no matter what the maximizing player does[†]. While a mixed strategy may guarantee a lower cost, the minmax strategy may be easier to determine and to implement. Also, it will often agree with the minimizing player's part of a pure strategy saddle point solution when such a solution exists.

Sufficient conditions for a minmax solution are presented for the case where the minimizing player uses feedback control and also where he uses open loop control. The results are illustrated with several examples.

2. FEEDBACK MINMAX STRATEGIES

Consider the differential system

$$\dot{x}(t) = f(x(t), u_1(t), u_2(t)) \tag{1}$$

where the state $x(t) \in R^n$, control variables $u_i(t) \in R^{m_i}$, $i = 1,2$. We assume $f(\cdot,\cdot,\cdot)$ is C^1 on $R^n \times R^{m_1} \times R^{m_2}$. The playing space is denoted by $X \times T$ where X and T are open sets in R^n and R^1 respectively. The game starts at a specified point $(x_o, t_o) \in X \times T$ and terminates at the smallest value of t_f, $t_o \le t_f < \infty$ for which $(x(t_f), t_f) \in \theta$, where the target set θ is a given set in the closure of $X \times T$, $\bar{X} \times \bar{T}$.

---

[1] The research was partially supported by ONR under Contract N00014-69-A-0200-1012.

[†] This statement will be made more precise in the next section.

For i = 1,2, let P_i denote the set of piecewise C^1 functions from X x T into R^{m_i}. A strategy $p_i(\cdot,\cdot)$ is admissible if, and only if, $p_i(\cdot,\cdot) \in P_i$ and $p_i(x,t) \in U_i$ for all $(x,t) \in X \times T$, where $U_i \subset R^{m_i}$ are given.

Let $N(p_1(\cdot,\cdot),x_o,t_o)$ denote the set of all admissible $p_2(\cdot,\cdot)$ such that $(p_1(\cdot,\cdot),p_2(\cdot,\cdot))$ is *playable at* (x_o,t_o), i.e. the pair $(p_1(\cdot,\cdot),p_2(\cdot,\cdot))$ generates a solution of (1) from (x_o,t_o), denoted by $x(\cdot)$ and called a trajectory, that reaches θ at some time $t_f \in \overline{T}$ and $x(t) \in X$ for all $t \in [t_o,t_f)$. Such a trajectory will be called a *terminating trajectory*. The set $N(p_1(\cdot,\cdot),x_o,t_o)$ is a subset of the admissible controls for player 2. In some cases, it may include all admissible $p_2(\cdot,\cdot)$. We assume that for all $p_1(\cdot,\cdot) \in P_1, N(p_1(\cdot,\cdot),x_o,t_o)$ is not empty.

For fixed $(x_o,t_o) \in X \times T$ and playable pair $(p_1(\cdot,\cdot),p_2(\cdot,\cdot))$, the cost is defined by

$$J(p_1(\cdot,\cdot),p_2(\cdot,\cdot),x_o,t_o) = \int_{t_o}^{t_f} L(x(t),p_1(x(t),t),p_2(x(t),t))dt \qquad (2)$$

where $L(\cdot,\cdot,\cdot)$ is C^1 on $R^n \times R^{m_1} \times R^{m_2}$. Note that the cost need not be unique since $(p_1(\cdot,\cdot),p_2(\cdot,\cdot))$ may generate more than one terminating trajectory.

Definition 1. Let $p_1^*(\cdot,\cdot)$ be admissible. Then $p_1^*(\cdot,\cdot)$ is a minmax strategy if, and only if, for all $(x_o,t_o) \in X \times T$

$$\max_{p_2(\cdot,\cdot) \in N(p_1^*(\cdot,\cdot),x_o t_o)} J(p_1^*(\cdot,\cdot).p_2(\cdot,\cdot),x_o,t_o)$$

is single-valued and

$$\max_{p_2(\cdot,\cdot) \in N(p_1^*(\cdot,\cdot),x_o t_o)} J(p_1^*(\cdot,\cdot),p_2(\cdot,\cdot),x_o,t_o)$$

$$\leq \max_{p_2(\cdot,\cdot) \in N(p_1(\cdot,\cdot),x_o t_o)} J(p_1(\cdot,\cdot),p_2(\cdot,\cdot),x_o,t_o)$$

for all admissible $p_1(\cdot,\cdot)$ and all terminating trajectories generated by $p_1(\cdot,\cdot)$ and the corresponding maximizing $p_2(\cdot,\cdot)$.

We now present a sufficient condition for minmax feedback control. Define

$$H(x,u_1,u_2,\lambda) = L(x,u_1,u_2) + \lambda f(x,u_1,u_2) \qquad (3)$$

where $\lambda \in R^n$.

Theorem 1. Let $p_1^*(\cdot,\cdot)$ be admissible. If there exists a $V(\cdot,\cdot): X \times T \to R^1$ that (α) is C^1 on X x T and (β) satisfies $V(x,t) = 0$ for all $(x,t) \in \theta$ and if there exists a piecewise C^1 function $\hat{u}_2(\cdot,\cdot,\cdot): X \times T \times U_1 \to U_2$ such that (γ) for all admissible $p_1(\cdot,\cdot), \hat{p}_2(\cdot,\cdot)$ is playable against $p_1(\cdot,\cdot)$ for every $(x_o,t_o) \in X \times T$ where $\hat{p}_2(x,t) = \hat{u}_2[x,t,p_1(x,t)]$ and (δ) for all $(x_o,t_o) \in X \times T$ the pair $[p_1^*(\cdot,\cdot),p_2^*(\cdot,\cdot)]$ leads to a unique terminating trajectory where $p_2^*(x,t) = \hat{u}_2[x,t,p_1^*(x,t)]$ and if for this $p_1^*(\cdot,\cdot)$, $V(\cdot,\cdot)$ and $\hat{u}_2(\cdot,\cdot,\cdot)$

(i) $V_t(x,t) + H[x,u_1,\hat{u}_2(x,t,u_1),V_x(x,t)] \geq 0$
for all $x \in X$, $u_1 \in U_1$ and almost all $t \in T$

(ii) $V_t(x,t) + H[x,p_1^*(x,t),p_2^*(x,t),V_x(x,t)] = 0$
for all $x \in X$ and almost all $t \in T$

(iii) $V_t(x,t) + H[x,p_1^*(x,t),u_2,V_x(x,t)] \leq 0$
for all $x \in X$, $u_2 \in U_2$ and almost all $t \in T$

then $p_1^*(\cdot,\cdot)$ is a feedback minmax strategy.

Proof: Consider any $(x_o,t_o) \in X \times T$. Let $p_1(\cdot,\cdot)$ be any admissible strategy and $x(\cdot)$ be a terminating trajectory corresponding to $[p_1(\cdot,\cdot),\hat{p}_2(\cdot,\cdot)]$ satisfying $x(t_o) = x_o$ with t_f the terminal time. For $[p_1^*(\cdot,\cdot),p_2^*(\cdot,\cdot)]$, the terminating trajectory will be denoted by $x^*(\cdot)$ and the terminal time by t_f^*. Note that

$$\frac{dV}{dt}(x(t),t) = V_t(x(t),t) + V_x(x(t),t)f[x(t),p_1(x(t),t),\hat{p}_2(x(t),t)] \tag{4}$$

Define the integral

$$I_1 \triangleq \int_{t_o}^{t_f} \frac{dV}{dt}(x(t),t)dt \tag{5}$$

From (α) and (β) of the Theorem

$$I_1 = -V(x_o,t_o) \tag{6}$$

and is invariant with the choice of $p_1(\cdot,\cdot)$. From (ii) and (4)

$$\int_{t_o}^{t_f} L[x^*(t),p_1^*(x^*(t),t),p_2^*(x^*(t),t)]dt = V(x_o,t_o) \tag{7}$$

Now consider

$$\Delta_1 \triangleq \int_{t_o}^{t_f} L[x(t),p_1(x(t),t),\hat{p}_2(x(t),t)]dt - \int_{t_o}^{t_f} L[x^*(t),p_1^*(x^*(t),t),p_2^*(x^*(t),t)]\,dt \tag{8}$$

Add $I_1 + V(x_o,t_o) = 0$ to (8). Then

$$\Delta_1 = \int_{t_o}^{t_f}\left\{\frac{dV}{dt}(x(t),t) + L[x(t),p_1(x(t),t),\hat{p}_2(x(t),t)]\right\} dt \tag{9}$$

and, from (i), $\Delta_1 \geq 0$. Using this and (8), we have

$$V(x_o,t_o) \leq \int_{t_o}^{t_f} L[x(t),p_1(x(t),t),\hat{p}_2(x(t),t)]dt \tag{10}$$

Thus,

$$V(x_o,t_o) \leq \max_{p_2(\cdot,\cdot) \in N(p_1(\cdot,\cdot),x_o,t_o)} \int_{t_o}^{t_f} L[x(t),p_1(x(t),t),p_2(x(t),)]\,dt \tag{11}$$

Since $p_1(\cdot,\cdot)$ was arbitrary, (11) holds for all admissible $p_1(\cdot,\cdot)$.

Let $\tilde{x}(\cdot)$ be a terminating trajectory corresponding to $[p_1^*(\cdot,\cdot),p_2(\cdot,\cdot)]$ where $p_2(\cdot,\cdot)\in N(p_1(\cdot,\cdot),x_o,t_o)$.

$$\Delta_2 \overset{\Delta}{=} \int_{t_o}^{t_f} L[\tilde{x}(t),p_1^*(\tilde{x}(t),t),p_2(\tilde{x}(t),t)]dt - V(x_o,t_o) \qquad (12)$$

Add $I_2 + V(x_o,t_o) = 0$ to (12) so that

$$\Delta_2 = \int_{t_o}^{t_f} \left\{\frac{dV_2}{dt}(\tilde{x}(t),t) + L(\tilde{x}(t),p_1^*(\tilde{x}(t),t),p_2(\tilde{x}(t),t))\right\} dt \qquad (13)$$

From (iii), $\Delta_2 \leq 0$ and from (12)

$$\int_{t_o}^{t_f} L(\tilde{x}(t),p_1^*(\tilde{x}(t),t),p_2(\tilde{x}(t),t))dt \leq V(x_o,t_o) \qquad (14)$$

Since (14) holds for any $p_2(\cdot,\cdot)\in N(p_1(\cdot,\cdot),x_o,t_o)$, (11) and (14) imply

$$\max_{p_2(\cdot,\cdot)\in N(p_1^*(\cdot,\cdot),x_o,t_o)} \int_{t_o}^{t_f} L[x(t),p_1^*(x(t),t),p_2(x(t),t)]dt$$

$$\leq \max_{p_2(\cdot,\cdot)\in N(p_1(\cdot,\cdot),x_o,t_o)} \int_{t_o}^{t_f} L[x(t),p_1(x(t),t),p_2(x(t),t)]dt$$

Since (x_o,t_o) was arbitrary, the theorem is proved. Q.E.D.

Comment 1. If H is separable, $H(x,u_1,u_2,\lambda) = H_1(x,u_1,\lambda) + H(x,u_2,\lambda)$, the theorem is quite similar to the standard sufficiency theorem for a pure strategy saddle point solution [5].

Comment 2. A method for selecting $\hat{u}_2(\cdot,\cdot,\cdot)$ which has proved successful is to choose it to satisfy

$$H[x,u_1,u_2,V_x(x,t)] \leq H[x,u_1,\hat{u}_2(x,t,u_1),V_x(x,t)]$$

for all $u_2\in U_2$.

Comment 3. From our experience with saddle point solutions, it is possible that $V(\cdot,\cdot)$ will not satisfy (α) of the theorem. The theorem can be extended to cover this situation by using the concept of a decomposition and only requiring $V(\cdot,\cdot)$ to be piecewise C^1.

Comment 4. It is conjectured that the necessary conditions which must be satisfied by a minmax strategy are the same as those for a pure strategy saddle point solution except that the generalized Weierstrass Condition or generalized Pontryagin Minimum Principle becomes

$$H[x^*(t),p_1^*(x^*(t),t),\hat{u}_2(x^*(t),t,p_1^*(x^*(t),t)),\lambda(t)]$$

$$\leq H[(x^*(t),u_1,\hat{u}_2(x^*(t),t,u_1),\lambda(t)]$$

for all $u_1\in U_1$ where $\lambda(\cdot)$ is the usual adjoint variable. Although this conjecture has not been proved, it can be used to generate candidates for a minmax strategy which can then be checked by the theorem.

Comment 5. If the cost contains a term of the type $\phi(x(t_f),t_f)$ then the theorem is applicable if (β) is modified to $V(x,t) = \phi(x,t)$ for all

$(x,t) \in \theta$.

3. AN EXAMPLE

A problem without a pure strategy saddle point solution, first presented by Berkovitz [2], is considered.

$$\dot{x}(t) = [u_2(t)-u_1(t)]^2$$

$$U_i \;\; = \{u_i(t) : 0 \le u_i(t) \le 1\}, \; i = 1,2$$

$$J(p_1(\cdot,\cdot),p_2(\cdot,\cdot),x_o,t_o) = \int_{t_o}^{t_f} x(t)dt$$

The playing space $X \times T = R^1 \times (-\infty,t_f)$ where t_f is specified. It will be shown that $p_1^*(x,t) = \frac{1}{2}$ for all $(x,t) \in X \times T$ is a minmax strategy.

Let $V(x,t) = \frac{1}{8}(t_f-t)^2 + x(t_f-t)$ and

$$\hat{u}_2(x,t,u_1) = \begin{cases} 1 & \text{if} \quad u_1 \ge \frac{1}{2} \\ 0 & \text{if} \quad u_1 > \frac{1}{2} \end{cases}$$

This $V(\cdot,\cdot)$ and $\hat{u}_2(\cdot,\cdot,\cdot)$, along with $p_1^*(x,t) = \frac{1}{2}$, satisfy (α)-(δ) of the Theorem and (i) - (iii). Thus $p_1^*(\cdot,\cdot)$ is a minmax feedback strategy.

If the cost J is replaced by -J and if $u_1(\cdot)$ and $u_2(\cdot)$ are interchanged, the theorem can be used to verify a maxmin strategy for Player 2. If this is done and $V(x,t) = -x(t_f-t)$ is used, it is easily verified that every admissible $p_2(\cdot,\cdot)$ is a maxmin strategy.

4. OPEN LOOP MINMAX CONTROL

The initial state (x_o,t_o) is now fixed. K_1 will denote the set of all piecewise continuous functions from T to R^{m_1}. A function $u_1(\cdot)$ is admissible if $u_1(\cdot) \in K_1$ and $u_1(t) \in U_1$ for all $t \in T$, where $U_1 \quad R^{m_1}$ is given. Admissible $p_2(\cdot,\cdot)$ are defined as before.

Let $\mathfrak{m}(u_1(\cdot))$ denote the set of admissible $p_2(\cdot,\cdot)$ such that $(u_1(\cdot), p_2(\cdot,\cdot))$ is playable at (x_o,t_o). It is assumed that for all admissible $u_1(\cdot)$, $\mathfrak{m}(u_1(\cdot))$ is not empty. For fixed (x_o,t_o), the cost is

$$J(u_1(\cdot),p_2(\cdot,\cdot)) = \int_{t_o}^{t_f} L(x(t),u_1(t),p_2(x(t),t))dt$$

Definition 2. An admissible $u_1^*(\cdot)$ is a minmax open loop control if, and only if,

$$\max_{p_2(\cdot,\cdot) \in \mathfrak{m}(u_1^*(\cdot))} J(u_1^*(\cdot),p_2(\cdot,\cdot))$$

is single-valued and

$$\max_{p_2(\cdot,\cdot) \in \mathfrak{m}(u_1^*(\cdot))} J(u_1^*(\cdot),p_2(\cdot,\cdot)) \le \max_{p_2(\cdot,\cdot) \in \mathfrak{m}(u_1(\cdot))} J(u_1(\cdot),p_2(\cdot,\cdot))$$

for all admissible $u_1(\cdot)$ and all terminating trajectories generated by $u_1(\cdot)$

and the corresponding maximizing $p_2(\cdot,\cdot)$.

Theorem 2. Consider a fixed $(x_o,t_o) \in X \times T$ and an admissible $u_1^*(\cdot)$. If there exist a $V_1(\cdot,\cdot) : X \times T \to R^1$ and a $V_2(\cdot,\cdot) : X \times T \to R^1$ that (α) are of class C^1 on $X \times T$ (β) satisfy $V_i(x,t) = 0$ for all $(x,t) \in \theta$, $i = 1,2$ (γ) satisfy $V_1(x_o,t_o) = V_2(x_o,t_o) \overset{\Delta}{=} V(x_o,t_o)$ and if there exists a piecewise C^1 function $\hat{u}_2(\cdot,\cdot,\cdot) : X \times T \times U_1 \to U_2$ such that (δ) for all admissible $u_1(\cdot)$, $\tilde{p}_2(\cdot,\cdot)$ is playable at (x_o,t_o) against $u_1(\cdot)$ where $\tilde{p}_2(x,t) = \hat{u}_2(x,t,u_1(t))$ (ϵ) $[u_1^*(\cdot),p_2^*(\cdot,\cdot)]$ leads to a unique terminating trajectory where $p_2^*(x,t) = \hat{u}_2(x,t,u_1^*(t))$ and if $V_1(\cdot,\cdot), V_2(\cdot,\cdot), u_1^*(\cdot)$ and $\hat{u}_2(\cdot,\cdot,\cdot)$ satisfy

i) $V_{1t}(x,t) + L(x,u_1,\hat{u}_2(x,t,u_1)] + V_{1x}(x,t)f[x,u_1,\hat{u}_2(x,t,u_1)] \geq 0$

$\forall\ x \in X$, $\forall\ u_1 \in U_1$ and for almost all $t \in T^o$

ii) $V_{1t}(x^*(t),t) + L(x^*(t),u_1^*(t),p_2^*(x^*(t),t))]$

$+ V_{1x}(x^*(t),t)f[x^*(t),u_1^*(t),p_2^*(x^*(t),t)] = 0$

for almost all $t \in T^o$

iii) $V_2(x,t) + L(x,u_1^*(t),u_2) + V_{2x}f(x,u_1^*(t),u_2) \leq 0$

$\forall\ x \in X, \forall\ u_2 \in U_2$ and for almost all $t \in T^o$ where $x^*(\cdot)$ is the terminating trajectory corresponding to $[u_1^*(\cdot),p^*(\cdot,\cdot)]$ and T^o is the subset of T with $t > t_o$, then $u_1^*(\cdot)$ is a minmax open loop control.

Proof: Since the proof is quite similar to the proof of Theorem 1, it is omitted.

Comment 6. If a single $V(\cdot,\cdot)$ can be found that satisfies (i), (ii) and (iii), this, of course, will verify sufficiency.

Comment 7. The theorem can also be proved if $V_2(\cdot,\cdot)$ replaces $V_1(\cdot,\cdot)$ in condition (ii).

Comment 8. A possible choice for $V_1(\cdot,\cdot)$ and $V_2(\cdot,\cdot)$, which has proved successful, is to choose V_1 to be the value of the closed loop game (assuming it exists) and V_2 to be the value of an optimal control problem formed by letting $u_1(t) = u_1^*(t)$ in the differential equations and cost.

Comment 9. Comments 2,3, and 5 also apply in the open loop case.

5. EXAMPLE 2

Consider the game with

$$\dot{x} = u_2 - u_1 \quad ; \quad x(t_o) = x_o$$

$$J(u_1(\cdot),p_2(\cdot,\cdot)) = -\frac{1}{2}x^2(t_f) + \frac{1}{2}\int_{t_o}^{t_f}\left[\frac{1}{3}u_1^2(t) - u_2^2(t)\right]dt$$

where (x_o,t_o) is prescribed, $X = R^1$, t_f is specified, $T = (t_o,t_f)$ and $U_1 = U_2 = R^1$.

It can be shown [6] that

$$3p_2^*(x,t) = p_1^*(x,t) = \frac{-3x}{1-2(t_f-t)}$$

is a closed loop saddle point solution if $t_f - t_o < \frac{1}{2}$ and if $t_f - t_o > \frac{1}{2}$ there is no closed loop saddle point solution. Also

$$3u_2^*(t) = u_1^*(t) = \frac{-3x_o}{1-2(t_f-t_o)}$$

is an open loop saddle point solution if $t_f - t_o < \frac{1}{3}$ and there is no open loop saddle point solution if $t_f - t_o > \frac{1}{3}$.

Now consider the game from Player 1's viewpoint when $t_f - t_o > \frac{1}{3}$ and Player 1 seeks a minmax open loop control. Theorem 2 will be used to show

$$u_1^*(t) = \frac{-3x_o}{1-2(t_f-t_o)}$$

is a minmax open loop control for Player 1 if $t_f - t_o < \frac{1}{2}$. Thus even though there is no open loop saddle point solution for $\frac{1}{3} < t_f - t_o < \frac{1}{2}$, Player 1 does have an open loop minmax control.

To prove the above statement, let

$$V_1(x,t) = -\frac{1}{2} Px^2, \quad V_2(x,t) = -\frac{1}{2} S_3x^2 - S_2xx_o - S_1x_o^2$$

where

$$\dot{P} + 2P^2 = 0, \quad P(t_f) = 1$$

$$\dot{S}_3 - S_3^2 = 0, \quad S_3(t_f) = 1$$

$$\dot{S}_2 - S_2S_3 + 3S_3P(t_o) = 0, \quad S_2(t_f) = 0$$

$$\dot{S}_1 - \frac{3}{2} P^2(t_o) - \frac{1}{2} S_2^2 + 3P(t_o)S_2 = 0, \quad S_1(t_f) = 0$$

so that

$$P(t) = \frac{1}{1-2(t_f-t)}, \quad S_3(t) = \frac{1}{1+t_f-t}$$

$$S_2(t) = \frac{3S_3(t)(t-t_f)}{1-2(t_f-t_o)}, \quad S_1 = \frac{-3S_2(t)(t-t_f)}{2[1-2(t_f-t_o)]} + \frac{3(t-t_f)}{2[1-2(t_f-t_o)]^2}$$

Conditions (α) and (β) are satisfied for $t_f - t_o < \frac{1}{2}$ while condition (γ) is verified in Appendix A of [7]. For $\hat{u}_2 = V_{1x} = -P(t)x$, (δ) and (ε) are verified using the basic existence theorem for linear systems, e.g. see Theorem 3-15 of [8].

Condition (i) is

$$-\frac{x^2}{2}\dot{P} + \frac{1}{6} u_1^2 - \frac{1}{2} V_{1x}^2 + V_{1x}(V_{1x} - u_1) \geq 0$$

for all $u_1 \in U_1$, $x \in R^1$ and almost all $t \in [t_o, t_f]$. Since $u_1 = 3V_{1x}$ is the absolute minimum of the left hand side and since it renders the left hand side zero, this condition is satisfied.

Condition (iii) is

$$-\frac{1}{2} \dot{S}_3x^2 - \dot{S}_2xx_o - \dot{S}_1x_o^2 + \frac{3}{2} x_o^2P^2(t_o) - \frac{1}{2} u_2^2 + V_{2x}(u_2 + 3x_oP(t_o)) \leq 0$$

for all $u_2 \in U_2$, $x \in R^1$ and almost all $t \in [t_o, t_f]$. Since $u_2 = V_{2x}$ maximizes the r.h.s and with this u_2 the r.h.s is zero, this condition is satisfied.

Condition (ii) is verified by noting that $x^*(t) = \theta(t,t_o)x_o$ where $\dot{\theta}(t,t_o) = 2P(t)\theta(t,t_o)$ and $\theta(t_o,t_o) = 1$ and using the fact that $\theta(t,t_o)P(t) = P(t_o)$. Thus (26) is a minmax control for Player 1 for $\frac{1}{3} < t_f - t_o < \frac{1}{2}$.

6. CONCLUDING REMARKS

One difficulty in using the theorems is verifying the playability condition (γ) of Theorem 1 or (δ) of Theorem 2. Most existence theorems for differential equations are local, rather than global, although some global results are available [1,9,10]. If the playability condition cannot be verified, some conclusions can still be reached. We will discuss these in terms of Theorem 1 and closed loop strategies, similar comments apply to the open loop situation.

If the minimizing player plays $p_1^*(\cdot,\cdot)$ and if the maximizing player must play a $p_2(\cdot,\cdot) \in N(p_1^*(\cdot,\cdot),x_o,t_o)$ then the minimizing player is guaranteed a cost of no more the $V(x_o,t_o)$. If (γ) is satisfied, then for any other $p_1(\cdot,\cdot)$, there is a $p_2(\cdot,\cdot) \in N(p_1(\cdot,\cdot),x_o,t_o)$ such that $J(p_1(\cdot,\cdot), p_2(\cdot,\cdot), x_o,t_o) \geq V(x_o,t_o)$. The minimizing player cannot guarantee improvement by choosing another $p_1(\cdot,\cdot)$. If (γ) is not satisfied, there may be a $\tilde{p}_1(\cdot,\cdot)$ that is not playable against $\hat{p}_2(\cdot,\cdot)$ which guarantees a cost $\tilde{V} < V(x_o,t_o)$ against all $p_2(\cdot,\cdot) \in N(\tilde{p}_1(\cdot,\cdot),x_o,t_o)$. In this case, $p_1^*(\cdot,\cdot)$ is a minmax strategy among all admissible $p_1(\cdot,\cdot)$ that are playable against $\hat{p}_2(\cdot,\cdot)$ rather than among all admissible $p_1(\cdot,\cdot)$. However, if this were the case, Player 2 could prevent Player 1 from guaranteeing $\tilde{V}$ by simply announcing he is going to play $p_2^*(\cdot,\cdot)$ which forces Player 1 to play $p_1^*(\cdot,\cdot)$. This leads to cost $V(x_o,t_o)$. The cost which can be guaranteed seems to depend on who announces his strategy first and whether or not none, one or both players must play so that termination occurs. Such questions cannot be answered in general, but only in specific problems.

REFERENCES

1. Friedman, A., Differential Games, Wiley-Interscience, 1971.
2. Berkovitz, L. D., "A Differential Game with No Pure Strategy Solution," Ann. Math. Stud., No. 52, Princeton University Press, Princeton, New Jersey, 1964, pp. 127-174.
3. Smoljakov, E. R., "Differential Games in Mixed Strategies," Soviet Math. Dokl. 11, 1970, pp. 330-334.
4. Elliot, R. J., Kalton, N. J., and Markus, L., "Saddle Points for Linear Differential Games," SIAM J. on Cont., Vol. 11, No. 1, 1973, pp. 100-112.
5. Leitmann, G., "Differential Games" in Differential Games: Theory and Applications, Ciletti, M. D. and Starr, A. W. ed., ASME 1970.
6. Schmitendorf, W. E., "Some Sufficiency Results for Differential Games," Paper presented at the Navy Workshop on Differential Games, Annapolis, Maryland, August 1973.
7. Schmitendorf, W. E., "Existence of Optimal Open-Loop Strategies for a Class of Differential Games," J. of Opt. Theory and Applic., Vol. 5, No. 5, May 1970.
8. Athans, M. and Falb, P. L., Optimal Control, McGraw Hill, 1966.
9. Coddington, E. A. and Levinson, N., Theory of Ordinary Differential Equations, McGraw Hill, 1955.
10. Birkhoff, B. and Rota, G., Ordinary Differential Equations, Blaisdell, 1969.

CONDITIONS FOR PARETO OPTIMALITY IN A MULTICRITERION SYSTEM WITH NONDIFFERENTIABLE FUNCTIONS.

JEAN-LOUIS GOFFIN
Assistant Professor
Ecole des Hautes Etudes Commerciales
Montréal, Québec, CANADA

Vector-valued optimization attempts to take into account directly the multiplicity of goals faced by decision-makers in engineering and economic systems.

The optimality concept most often used when the performance of a system is described by a vector-valued criterion is that of Pareto.

Characterizations of Pareto-optimum points can be obtained through two different approaches :

(i) the scalarization process
(ii) a direct approach.

The results of [1,3,5] can be extended to the case where the vector-valued criterion is not differentiable :

(i) the scalarization process does not depend on any differentiability assumptions, but some convexity properties are required in order to obtain necessary conditions for optimality (the vector-valued criterion must be γ-vex on the feasible set [5,6]).

(ii) the direct approach requires some weak differentiability conditions : if the objective functions represent costs (i.e. we minimize) their directional derivatives must be sublinear functionals [see 2] of the direction.

Assumptions :

1. $\varphi : R^n \rightarrow R^p : x \rightarrow \varphi(x) = (\varphi_j(x)) \quad j = 1,\ldots,p$

   the vector-valued cost is continuous.

2. $\varphi_j'(x,h) : R^n \times R^n \rightarrow R$

   the directional derivative of φ_j is a sublinear function of h.

3. The directional derivative satisfies :

$$\lim_{h\to 0} \frac{\varphi_j(x+h) - \varphi_j(x) - \varphi_j'(x,h)}{||h||} = 0$$

This research was supported by le Ministère de l'Education Nationale du Québec (D.G.E.S.).

4. $X \subset R^n$

the feasible set is defined by a set of constraints :

$X = \{x \in R^n : g_i(x) \geq 0, \ i = 1,\ldots,m\}$

$g_i = R^n \to R$ is C^1 $\quad i = 1,\ldots,m,$

the gradient of g_i is denoted by g_i'.

In addition the constraints satisfy a constraint qualification (Kuhn-Tucker, etc...).

Theorem :

Under assumptions (1,2,3,4) if x^* in X is Pareto-Optimal, then there exist :

$\lambda_i \geq 0, \ i = 1,\ldots,m$

and :

$\alpha_j \geq 0, \ j = 1,\ldots,p, \ \alpha_l > 0$ for some l,

such that :

$$\sum_{j=1}^{p} \alpha_j \varphi_j'(x^*,h) \geq \sum_{i=1}^{m} \lambda_i < g_i'(x^*), h > \qquad (1)$$

and :

$$\lambda_i g_i(x^*) = 0 \qquad i = 1,\ldots,m. \qquad (2)$$

Furthermore, if the g_i's are concave and the φ_j's convex, then (1) and (2) with :

$\lambda_i \geq 0, \qquad i = 1,\ldots,m$

and :

$\alpha_j > 0, \qquad j = 1,\ldots,p$

implies that x^* is Pareto Optimal.

Remark :

Assumptions (1,2,3) are satisfied in the following cases :

(i) each φ_j is the sum of a convex function and a differentiable function.

(ii) $\varphi_j(x) = \underset{y \in Y_j}{\text{Max}} \ \psi_j(x,y)$

where Y_j is a compact space, ψ_j is continuous in x and y, and $\nabla_x \psi_j(x,y)$ is continuous in x and y.

REFERENCES

[1] N.O. da CUNHA & E. POLAK :

"Constrained Minimization under Vector-Valued Criteria in Linear Topological Spaces", in "Mathematical theory of control". Academic Press (BALAKRISHNAN, NEUSTADT editors). (1967).

[2] Vladimir F. DEMYANOV & Aleksandr M. RUBINOV

"Approximate Methods in Optimization Problems". Elsevier Publishing Company, Inc. New York (1970).

[3] Alain HAURIE & Jean-Louis GOFFIN

"Necessary Conditions and Sufficient Conditions for Pareto Optimality in a Multicriterion Perturbed System". Proceedings of the tenth Annual Allerton Conference on Circuit and System Theory. (1973).

[4] Alain HAURIE

"On Pareto Optimal Decisions for a Coalition of a Subset of Players", I.E.E.E. Trans. on Automatic Control, April 1973.

[5] Jean-Louis GOFFIN & Alain HAURIE

"Necessary Conditions and Sufficient Conditions for Pareto Optimality in a Multicriterion Perturbed System". To appear in the Proceedings of 5th IFIP Conference on Optimization Techniques. (1973).

[6] J.P. AUBIN

"Existence of Saddle Points for Classes of Convex and Non-Convex Functions". To appear in the Proceedings of 5th IFIP Conference on Optimization Techniques. (1973).

COST-EFFECTIVE DESIGN OF SPECIAL-PURPOSE PROCESSORS: A FAST FOURIER TRANSFORM CASE STUDY

A. G. Larson
Radio Physics Laboratory
Stanford Research Institute
Menlo Park, California 94025

E. S. Davidson
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

ABSTRACT

A design methodology is described for implementation, in hardware, of a computation to meet performance and cost criteria that can maximize the ratio of processing performance to processor cost. Potential designs permit consideration of alternative computation algorithms, circuit technologies, and processing architectures.

The fast Fourier transform (FFT) computation is used as a case study to illustrate several practical considerations. Several new pipelined cascade implementations of the FFT offer highest performance/cost ratio.

INTRODUCTION

The digital system designer must select an appropriate algorithm, a desirable circuit technology, and a suitable processing architecture for the implementation of a particular computational sequence to produce a cost-effective design within acceptable bounds of processing time and cost. Often, to achieve certain performance or cost goals, he must consider specialized processing organizations having local processing architectures that are suitable for each possible algorithm and compatible with the global system environment. This paper examines alternative architectures employing pipelined and parallel structures for a subsystem designed to perform a particular well-defined class of computations using given algorithms and circuit technologies. Comparisons may also be made across alternative algorithms and technologies.

For our purposes, performance is derived from aspects of the subsystem architecture and thus represents the potential performance of that subsystem. Cost, likewise, is subsystem cost. Actual system cost involves adding a control cost, dependent on the system in which the subsystem is embedded. However, the fractional increase of subsystem cost may be much greater than the fractional increase in total system cost, unless extensive total system modification is required to accommodate an enhanced subsystem. In the design strategy proposed here, local performance and cost of alternative architectures are analyzed first, and a locally cost-effective design is generated. This design is then modified if necessary in view of global performance and cost.

In this paper the procedures for cost-effective logic design are applied to the fast Fourier transform (FFT) algorithm as an extended example of a computational sequence that is widely used in a well-known class of special-purpose processing hardware. The FFT is a useful example for contrasting alternative processing architectures, since it exhibits both a highly regular, but nontrivial, replication pattern, and a great degree of independence or parallelism (as well as much dependence) among its component steps [1]. The FFT is also of great practical significance, and several FFT hardware implementations exist employing various degrees of pipelining and parallelism for performance enhancement [2-4].

SIX ARCHITECTURES

Fundamental to the development of a cost-effective design methodology are two basic architectures: simple and cascade. Each of these is enhanced through parallelism to parallel simple sections and parallel cascade sections. The cascade architecture may be enhanced through pipelining to pipelined cascade and parallel pipelined cascades. The relevant local characteristics of these six architectures, in terms of performance, cost, and the ratio of performance to cost, are presented in Table I.

The simple architecture has one computation element, S. It delivers its output to a register, R. The costs of these components are also called S and R, respectively. The times required by S and R are τ and ρ, respectively, which are defined so that the simple architecture can perform a new computation each $\tau + \rho$ seconds.

The cascade architecture has k computational elements, S_1, S_2, ..., S_k, and an output register, R. These elements cost an average of $\bar{S}$ each and require an average time of $\bar{\tau}$ each. The cascade architecture can produce one result from each element in $k\bar{\tau} + \rho$ seconds, giving a computation rate of k elementary results per $k\bar{\tau} + \rho$ seconds.

Two parallel architectures are based on the simple and cascade architectures. In parallel simple architecture, there are n parallel sections each of which has simple architecture. Thus, the cost is n times the cost of a simple section, and n elementary results can be produced each $\tau + \rho$ seconds. In parallel cascade architecture, each of the n parallel sections is a cascade of k simple elements. The parallel cascade thus has a cost of $n(k\bar{S} + R)$ and can produce nk elementary results each $k\bar{\tau} + \rho$ seconds. These parallel architectures thus have a cost of n times the section cost and a performance of n times the section performance. The performance/cost ratio is thus not a function of n.

The final two architectures are pipelined architectures. The cascade architecture may be pipelined by adding a register between each adjacent pair of segments in the cascade, resulting in the pipelined cascade architecture. It is usually desirable to clock the registers synchronously. The clock has period $\hat{\tau}$ = max (time delay of S_i). The delay of the cascade is then $k(\hat{\tau} + \rho)$. Its cost is $k(\bar{S} + \bar{R})$, where $\bar{R}$ is the cost of the average register. Note that the various registers may not have equal numbers of bits or equal cost. The cascade should be designed to have minimum $\bar{R}$ as well as to have $\hat{\tau}$ as close to $\bar{\tau}$ as possible. These two objectives cannot always be met simultaneously.

The synchronization and independence of computational elements provided by the added registers allows a new input set to be introduced and a new result obtained each $\hat{\tau} + \rho$ seconds. Each output result passes through the entire cascade and thus represents k elementary results. Thus in $k(\hat{\tau} + \rho)$ seconds, k^2 elementary results can be produced. For parallel pipelined cascades, the performance relative to a single pipelined cascade is n times as high, as is the cost. The performance/cost ratio is once again independent of n.

A simple architecture (taken as a standard throughout) can be converted into a cascade by successive replications of the simple element in cascade, or by subdivision (a cascade decomposition) of the simple element. A cascade can be pipelined, and any of these three architectures can be replicated in parallel. These relations produce ten alternative realizations based on a simple element, summarized Table II, with but three distinct performance/cost formulae.

We have used the notation k_c and k_p to indicate the distinction between the number of elements within a cascade architecture formed by replication and the number of elements within a pipelined cascade architecture formed by subdivision, respectively. Also listed are the assumptions made to compare performance/cost among the six architectures. Let us define the performance/cost ratio of a particular architecture as its value, or figure of merit, M. We may represent the distinct performance/cost expressions of Table II for simple architecture, cascade by replication, and pipelined cascade by subdivision by M_s, $M_c(k_c)$, and $M_p(k_p)$ respectively. Note that $M_s = M_c(k_c) = M_p(k_p)$ for $k_c = k_p = 1$. Furthermore, it may be seen that $M_p(k_p) = M_c(1/k_p)$. Thus, only $M_p(k)$ is necessary to characterize all three formulae, where $k \geq 2$ for integer k is realizable in a pipelined cascade achieved by subdivision of a simple element into k subelements, $k = 1$ is realizable in a simple architecture, and $k = 1/i$ for integer $i \geq 2$ is realizable in a cascade architecture achieved by replication of a simple element i times.

A METHODOLOGY FOR MAXIMUM-VALUE DESIGN

In this section we summarize a logic design procedure that is presented in greater detail elsewhere [5,6]. It is assumed in the following that a given computational algorithm will be implemented using a circuit technology for which the delay and cost of functional arithmetic and register modules are known.

It may be shown that the value of k, possibly non-integer, that optimizes circuit value (locally defined performance per unit processor cost) is

$$k_{opt} = \sqrt{\frac{\tau S}{\rho R}} .$$

If $k_{opt} \leq 1$, the maximum-value design consists of $1/k_{opt}$ replications of the simple element in cascade, while if $k_{opt} > 1$, the maximum-value design is a pipelined cascade achieved by subdivision of the simple element into k_{opt} subelements [6].

Designs utilizing any of the common medium-scale-integration (MSI) circuit technologies for implementation of any reasonably complex processing structure will have both $S > R$ and $\tau > \rho$; thus, $k_{opt} > 1$. For example, the architecture for cost-effective design of any FFT processing structure implemented with these MSI components will then be pipelined cascade formed by subdivision of an FFT simple element.

As an initial, if impractical, design, consider a simple element that performs the entire algorithm to be implemented. If necessary, this element may be made combinational by expanding sequential parts of the logic into iterative arrays of next state/output logic. We begin by stripping away parallel and/or cascade-by-replication features of the implementation. This procedure continues until the resultant structure is not further reducible or until the current value of k_{opt} is less than unity.

Thus if the simple one-pass implementation of the algorithm consists entirely of n parallel copies of a section of the design, only one copy is retained in a new implementation. If, at this point, the new implementation has $k_{opt} \leq 1$, it is accepted as the maximum-value implementation since the computation cannot make use of cascade replications of the simple element.

For the case $k_{opt} > 1$, cascade replications within the element are removed. If $k_{opt} < 1$ after removing cascade replication and parallelism, some cascade-by-replication designs should be considered to restore some degree of replication. High-local-value designs have k_c close to $1/k_{opt}$, whereas for efficiently utilizable designs, k_c should evenly divide r, the degree of replication. Note that $r > 1/k_{opt}$, since using $k_c = r$ restores the original degree of replication in cascade for which $k_{opt} \geq 1$, indicating that more replication was present than desirable.

Finally, if, after replication and parallelism are removed, it still remains true that $k_{opt} \geq 1$, then a pipelined cascade by subdivision must be employed to yield the maximum-value implementation, since removing more replication or parallelism is undesirable or impossible. It would be expected that this would be the case in any reasonably complex computational sequence (such as the FFT). Here, k_{opt} will indicate the degree of subdivision for a maximum-value pipelined cascade. Several practical considerations can arise that influence the appropriate number of elements in the pipeline.

The k_p chosen must be integer and should be close to k_{opt} for maximum value. $\bar{R}$ for k_p may not be equal to R in that $\bar{R}$ may be a function of k_p. For certain logic-circuit technologies -- e.g., noncustom MSI, cut sets may not occur freely but must avoid passing within circuit elements. Also, in some technologies $\hat{\tau}$ may not be τ/k_p or even very close if, for example, the delay of several levels of logic is substantially less than the number of levels times a single gate delay. Each of these considerations may cause an adjustment in the k_p selected for maximum value.

When the potential performance capability of the maximum-value design is too low, a new architecture must be found based on the maximum-value implementation that will not lower the value of the implementation unnecessarily. For small increase in performance, using $k > k_{opt}$ may achieve the necessary performance at least cost. For improvements in performance by a factor of 2 or more, the architecture that will accomplish this goal is parallel pipelined cascade by replication, with the maximum-value design as a simple element. The design involves degree n parallelism and k replications in cascade, where kn is the performance-improvement factor desired. The factor kn also represents the cost-increase factor over the previous maximum-value implementation, so that no degradation in value results. Naturally either k or n can be held to unity if desirable. Utilizing $k > 1$ depends on as yet unused replications in cascade within the computation. In addition, excess parallelism of degree n must be available in the computation or in a parallel or interleaved set of computations in order to utilize the increased performance. In the ideal case, the effect of utilizing this implementation is to increase design value of the final subsystem above the value obtained by the more common practice of utilizing some mix of parallelism and pipelined cascades by replication based on a nonmaximum-value simple element. The maximum increase is $\frac{k_{opt}}{4} + \frac{1}{2} + \frac{1}{4\,k_{opt}}$, where k_{opt} is $\sqrt{\frac{\tau}{\rho}\frac{S}{R}}$ [6].

It should be stated again that this design methodology is carried out on a local basis, under certain assumptions. The major assumptions are that delay is a linear function of the number of levels of combinational logic and, for pipelining by subdivision, $\bar{R} = R$, $\hat{\tau} = \tau/k$, and ρ is constant. Of course, only integer or reciprocal integer values of k are meaningful. In general, $\bar{R} \geq R$ and $\hat{\tau} > \tau/k$, although $\bar{R} < R$ is possible if small internal cuts sets can be found or if some latches may be used internally.

Register delay ρ is generally constant, but may nearly vanish in some cases when the design permits latches to be used internally.

Values of ρ and $\bar{R}$ that are smaller than those assumed tend to increase the attractiveness of pipelining to achieve best performance/cost ratios, and tend to increase k_{opt} as well; larger values of $\hat{\tau}$ and $\bar{R}$ than those assumed tend to lower pipelining performance/cost ratios and decrease k_{opt}. Each of these practical and technology and implementation-dependent considerations may limit the maximum attainable circuit value, but these considerations can be incorporated into a more detailed and accurate design for a particular application environment.

COST-EFFECTIVE DESIGN OF THE FFT BUTTERFLY

The radix-2 FFT algorithm computation structure has $\frac{N}{2} \log_2 N$ complex computation operations organized in $\log_2 N$ stages of $\frac{N}{2}$ parallel computations, where the transform size N is an integer power of two [1]. The first step in the evolution of a cost-effective design of an FFT processor is the development of a single fully combinational processing element to perform the entire FFT algorithm in one pass. This processing element is then decomposed by application of cost-effective design procedures [6].

Under this decomposition, all parallel and cascade-by-replication features are stripped away until the resultant structure is not further reducible. Since the overall FFT-processing element consists of $\log_2 N$ cascade replications of a single transform stage, we first eliminate $(\log_2 N) - 1$ copies of this stage. The remaining transform state consists of $\frac{N}{2}$ parallel replications of one section of its design. We then eliminate $\frac{N}{2} - 1$ copies of this section. This will reduce the original processing element to a two-point complex transform, known as the FFT butterfly transform.

The decimation-in-time FFT butterfly algorithm has within it one further replication, of the real or imaginary component calculation, that may also be eliminated [6]. This resultant computation structure is denoted as the FFT simple element.

The FFT simple element is then viewed as a cascade of subelements, each specified by a known logic delay and cost for a given circuit technology. For convenience of illustration, the FFT processor designs are constrained to use noncustom MSI arithmetic-logic and register modules as representative of the most common, yet flexible, circuit technology in current use. These designs employ commercially available transistor-transistor-logic circuit modules that were chosen to span typical speed and cost characteristics of current MSI components, ranging from relatively low-speed and inexpensive computation modules and registers to their relatively high-speed and expensive equivalents within the Schottky family.

Two FFT butterfly computation algorithms are considered in this study. The first is a standard FFT half-butterfly implementation, requiring both a multiply operation and a set of trigonometric coefficients to perform a complex vector rotation. The necessity for both an explicit multiply operation and any stored trigonometric coefficients, and all associated costs, can be avoided, however, by use of a balanced cordic algorithm in implementing the butterfly transform calculations [7]. The second FFT butterfly design makes use of this technique in a cordic FFT full-butterfly implementation.

For any combination of MSI arithmetic and register modules considered to realize either simple-element implementation, it is found that $k_{opt} \gg 1$. Thus, cost-effective FFT processors implemented with these modules must be based on a pipelined cascade architecture formed by subdivision of the FFT butterfly element.

The number of segments in the pipeline will be determined not only by the technology-dependent factors, and the simple-element-computation algorithm, but by several practical considerations as well. The most dominant of these in actual circuit designs is the constraint on selection of the optimum cut set to avoid passing within individual MSI modules. Not only does this constraint place a physical bound, K, on k_p, which may be less than the desired k_{opt}, but for k_p near K, undesirable performance bottlenecks may result.

A pipeline segment producing a $\hat{\tau}$ much greater than τ/k_p is called a performance bottleneck. If the best k_p-stage cut set contains a performance bottleneck, the subsystem designer must also consider parallel replication of the segment causing the unacceptable $\hat{\tau}$ as a means of reducing $\hat{\tau}$ to be closer to τ/k_p [8]. Parallel replication of the bottleneck should be considered not only for improvement of processing performance, but also for a potential increase of the simple-element implementation value. In application, this technique will require the additional time delay and cost of a multiplexing element to cycle the pipeline data flow alternately between the two copies of the replicated segment. It is often found that multiplexing between parallel replications of a segment can reduce the time delay of that segment by a factor of nearly two, for somewhat greater than twice the segment cost. When a single segment limits performance but does not dominate the cost of a pipelined cascade, a significant improvement in design value can be attained through multiplexing. The additional delay and cost of the multiplexer component must, of course, be considered. It is found in practice that multiplexer costs approach those of the pipeline segment itself. Thus, parallel replication of the bottleneck segment may not always improve the design value, but should nevertheless be investigated for each circuit implementation.

It is also essential in all FFT processor implementations, of course, that the bandwidth of external memory be sufficient to meet the desired data-processing bandwidth. This important, and often limiting, factor is considered a system-embedding detail that is beyond the scope of our local design; we then further assume that the attainable FFT-processing bandwidths of our designs are not bound by considerations of data flow or storage. In light of this limitation, however, a desired processing bandwidth for a particular application may be met.

FFT BUTTERFLY DESIGN RESULTS

In this section we present the results of a design study for the cost-effective implementation of an FFT processor, using each of two different FFT simple-element-implementation algorithms and a pipelined cascade-by-subdivision processor architecture. A detailed description of the overall design parameters and logic organization is given elsewhere [6].

Design emphasis is placed on the development of maximum-value designs and the performance enhancement of simple FFT architectures at minimum cost. When necessary to further improve the design value, pipeline segments causing performance degradation due to excessive computation-time delay are duplicated in parallel, with computation flow multiplexed between them. For each computation algorithm a wide range of alternative designs is also developed that may be useful for meeting other design goals. Maximum-performance and minimum-cost designs are specifically indicated.

Figure 1 illustrates typical design results for various standard half-butterfly design goals. The computed design value is plotted as a function of the number of pipeline stages used in each particular implementation. The computation structure for each implementation consists of seven arithmetic cascade segments, or $K = 7$. A theoretical maximum-value design is also calculated for this implementation, as shown in Figure 1, attaining a peak value at $k_p = k_{opt} = 8$. The actual maximum implementation value, occurring at $k_p = 5$, is approximately 13 percent lower than the peak estimated design value at k_{opt}.

The lowest-cost design and the maximum-performance design are also shown for comparison to the maximum-value design. It is noted that the lowest-cost design is also the best-value design for a k_p of 3 or 4 stages. The maximum-performance design, making full use of Schottky components, is seen to have a relatively stable $M_p(k_p)$ for $k_p \geq 3$, although the calculated k_{opt} is 11 pipeline stages. The value of the maximum-performance design is actually closer, in general, to the design with the lowest attained value than to the design with the highest value over all processor designs for each number of pipeline stages.

The balanced cordic FFT full-butterfly implementation consists of twenty arithmetic cascade segments organized in diagonal slices across the balanced cordic computation structure. Figure 2 illustrates typical design results for various design performance and cost goals. It is noted that in any of these balanced cordic algorithm implementations, the technique of parallel replication of cascade segments to eliminate a bottleneck $\hat{\tau}$ is not relevant since the time delay per segment is nearly constant.

The calculated k_{opt} for the maximum-value design is 24 pipeline stages. As with the standard butterfly design, we are again able to come relatively near the theoretical optimum number of pipeline stages in our maximum-value design. At $k_p = 20$, the actual implementation value of the maximum-value design is approximately 10 percent lower than the peak estimated design value at k_{opt}. The theoretically derived peak implementation value is thus shown to be a useful estimator of the highest $M(k_p)$ attainable in actual designs, due to the relative flatness of the $M(k_p)$ function for k_p near k_{opt} even when the optimal k_p cannot be physically realized and the assumptions that $\hat{\tau} = \tau/k_p$ and $\bar{R} = R$ are not met exactly.

The maximum-value-design components produce the highest-value implementation for all k_p. It is again seen that the value of the maximum-performance design is closer to the design with the lowest attained value than to the design with the highest attained value. One may also note that even with the poorest choice of circuit-components, pipelining the balanced-cordic-algorithm design can attain higher value than any of the nonpipelined implementations. Thus, the effect of using a pipelined architecture is seen to be of greater design-value significance than the choice of circuit technology.

The absolute performance and cost for these various FFT implementations are shown in Figure 3, with performance expressed in terms of full-butterfly transforms completed per second and with cost expressed as the actual component cost of each implementation. Five design goals are illustrated for FFT processors using both the standard half-butterfly and the balanced cordic full-butterfly computation algorithms. The performance and cost achieved by replication in parallel of each maximum-value design are also displayed.

The actual potential of cost-effective design is best seen in comparison to current FFT-processor design practice. In the case of either the standard or balanced cordic butterfly designs, the maximum-value implementations are of significantly greater design value than are the nonpipelined designs (cases 2, 4, 7, and 9 in Figure 3). It may be seen in Figure 3 that an FFT processor implemented with replications of either maximum-value-design pipelined butterfly computation element can provide an FFT-processing bandwidth equal to that of several nonpipelined butterfly computation elements (representative of prevalent high-performance FFT-implementation practice) at 1/5 to 1/2 the overall component cost.

These cost-effective design results specifically pertain to FFT processors; however, they are also general in that they are indicative of the design techniques that may be applied to develop processing subsystems to implement any similar calculation sequence. For FFT processors, this design technique is, of course, not restricted to the implementation of any particular radix-2 FFT-algorithm processor. The methodology may be directly extended to the implementation of the exact transform-calculation sequence necessary for arbitrary transform size.

ACKNOWLEDGMENT

This research was supported by Stanford Research Institute, Menlo Park, California, and by National Science Foundation Grant GJ-35584X.

REFERENCES

[1] W. T. Cochran et al., "What is the Fast Fourier Transform?," IEEE Trans. Audio Electroacoust., Vol. AU-15, No. 2, pp. 45-55, June 1967.

[2] G. D. Bergland, "Fast Fourier Transform Hardware Implementations -- An Overview," IEEE Trans. Audio Electroacoust., Vol. AU-17, No. 2, pp. 104-108, June 1969.

[3] H. L. Groginsky and G. A. Works, "A Pipeline Fast Fourier Transform," IEEE Trans. Computers, Vol. C-19, No. 11, pp. 1015-1019, November 1970.

[4] B. Gold and T. Bially, "Parallelism in Fast Fourier Transform Hardware," IEEE Trans. Audio Electroacoust., Vol. AU-21, No. 1, pp. 5-16, February 1973.

[5] E. S. Davidson and A. G. Larson, "Pipelining and Parallelism in Cost-Effective Processor Design," submitted for publication.

[6] A. G. Larson, "Cost-Effective Processor Design with an Application to Fast Fourier Transform Computers," Ph.D. Dissertation, Digital Systems Laboratory Technical Report 73, Stanford University, Stanford, California, August 1973.

[7] A. M. Despain, "Fourier Transform Computers Using CORDIC Iterations," submitted for publication.

[8] L. E. Shar, "Design and Scheduling of Statically Configured Pipelines," Ph.D. Dissertation, Digital Systems Laboratory Technical Report 42, Stanford University, Stanford, California, August 1972.

Table I

PARAMETERS OF ALTERNATIVE ARCHITECTURES

Architecture	No. of Parallel Sections	No. of Elements per Section	Calculation Time	Elementary Results	Cost	Performance/Cost (Elementary Results/Time/Cost)
Simple	1	1	$\tau + \rho$	1	$S + R$	$\frac{1}{(S + R)(\tau + \rho)}$
Cascade	1	k	$k\bar{\tau} + \rho$	k	$k\bar{S} + R$	$\frac{k}{(k\bar{S} + R)(k\bar{\tau} + \rho)}$
Parallel Simple Sections	n	1	$\tau + \rho$	n	$n(S + R)$	$\frac{1}{(S + R)(\tau + \rho)}$
Parallel Cascade Sections	n	k	$k\bar{\tau} + \rho$	nk	$n(k\bar{S} + R)$	$\frac{k}{(k\bar{S} + R)(k\bar{\tau} + \rho)}$
Pipelined Cascade	1	k	$k(\hat{\tau} + \rho)$	k^2	$k(\bar{S} + \bar{R})$	$\frac{1}{(\bar{S} + \bar{R})(\hat{\tau} + \rho)}$
Parallel Pipelined Cascades	n	k	$k(\hat{\tau} + \rho)$	nk^2	$nk(\bar{S} + \bar{R})$	$\frac{1}{(\bar{S} + \bar{R})(\hat{\tau} + \rho)}$

Table II

PERFORMANCE/COST COMPARISON OF ARCHITECTURES RELATED BY REPLICATION OR SUBDIVISION

Architecture	Performance/Cost (Simple Results/Time/Cost)	
Simple	$\frac{1}{(S + R)(\tau + \rho)}$	
Parallel Simple Sections	$\frac{1}{(S + R)(\tau + \rho)}$	
	Replication	Subdivision
	$[\bar{S} = S;\ \bar{\tau} = \tau = \hat{\tau};\ \bar{R} = R]$	$\Big[\bar{S} = \frac{S}{k};\ \bar{\tau} = \frac{\tau}{k} = \hat{\tau};\ \bar{R} = R;$ k elementary results = 1 simple result$\Big]$
Cascade	$\frac{k_c}{\left(k_c S + R\right)\left(k_c \tau + \rho\right)}$	$\frac{1}{(S + R)(\tau + \rho)}$
Parallel Cascade Sections	$\frac{k_c}{\left(k_c S + R\right)\left(k_c \tau + \rho\right)}$	$\frac{1}{(S + R)(\tau + \rho)}$
Pipelined Cascade	$\frac{1}{(S + R)(\tau + \rho)}$	$\frac{k_p}{\left(S + k_p R\right)\left(\tau + k_p \rho\right)}$
Parallel Pipelined Cascades	$\frac{1}{(S + R)(\tau + \rho)}$	$\frac{k_p}{\left(S + k_p R\right)\left(\tau + k_p \rho\right)}$

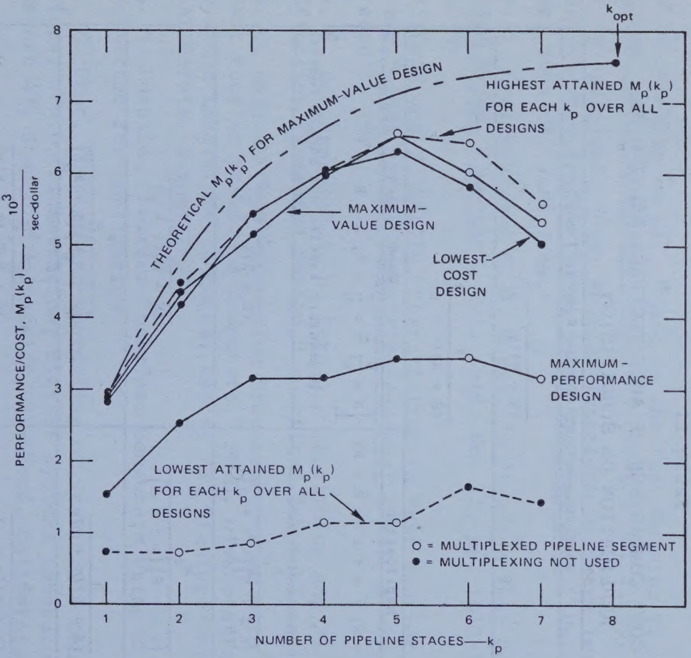


Figure 1 Design Value for Various Standard Half-Butterfly Design Goals

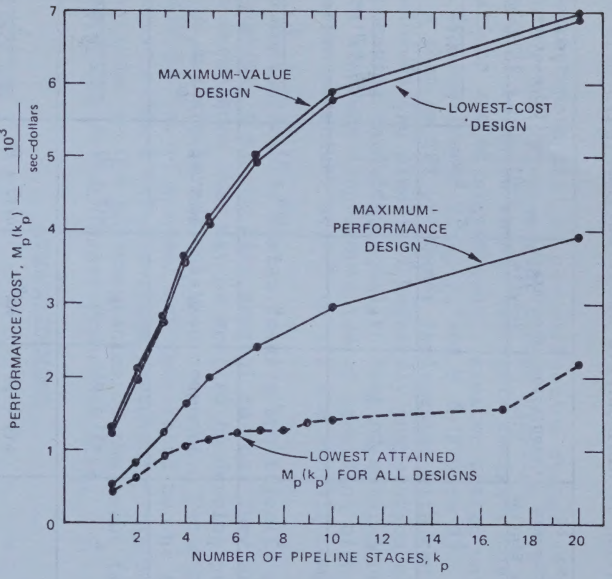


Figure 2 Design Value for Various Balanced Cordic Full-Butterfly Design Goals

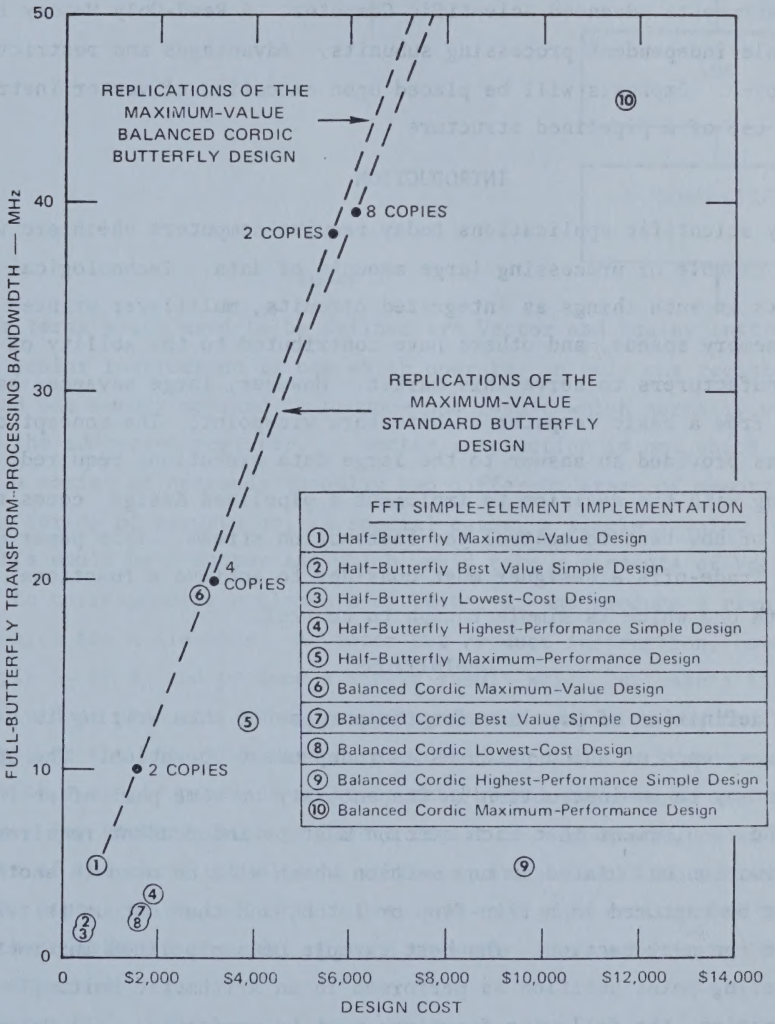


Figure 3 Performance and Cost for Several FFT Implementations

CONTROL OF A VARIABLE CONFIGURATION PIPELINED ARITHMETIC UNIT

CHARLES M. STEPHENSON
Texas Instruments Incorporated
Austin, Texas

ABSTRACT

This paper discusses the control aspects of the Arithmetic Unit of the Texas Instruments Advanced Scientific Computer. A Read-Only Memory is used to schedule independent processing subunits. Advantages and restrictions are explored. Emphasis will be placed upon execution of vector instructions by use of a pipelined structure.

INTRODUCTION

Many scientific applications today require computers which are very fast and capable of processing large amounts of data. Technological developments in such things as integrated circuits, multilayer printed circuit boards, memory speeds, and others have contributed to the ability of computer manufacturers to serve this market. However, large advances had to be realized from a basic computer architecture viewpoint. The concept of pipelining has provided an answer to the large data executions required.

Along with the decision to implement a pipelined design, comes the question of how best to control the instruction stream. This paper deals with the trade-offs a designer must consider to achieve a functional design, along with one which is simple enough to control.

DEFINITIONS

The definition of pipeline for the purpose of this writing is a series of sections, each of which performs an independent operation. The operation performed may be an instruction in its entirety or some part of an instruction. The requirement that each section must be independent requires that any information calculated in one section which will be used in another section must be captured in a flip-flop or latch, and that output is transmitted to the next section. The best example of a pipelined instruction may be a floating-point addition as performed in an Arithmetic Unit. To execute this operation, the following functions must be performed: (1) Determine the difference in the two exponents, (2) Shift the smaller operand to the right to align the fractions, (3) Add the fractions, and (4) Normalize the result. This instruction could be performed with either 1, 2, 3, or 4 pipe sections, depending upon the goal of the designer. Three possible pipelines are shown in Figure 1.

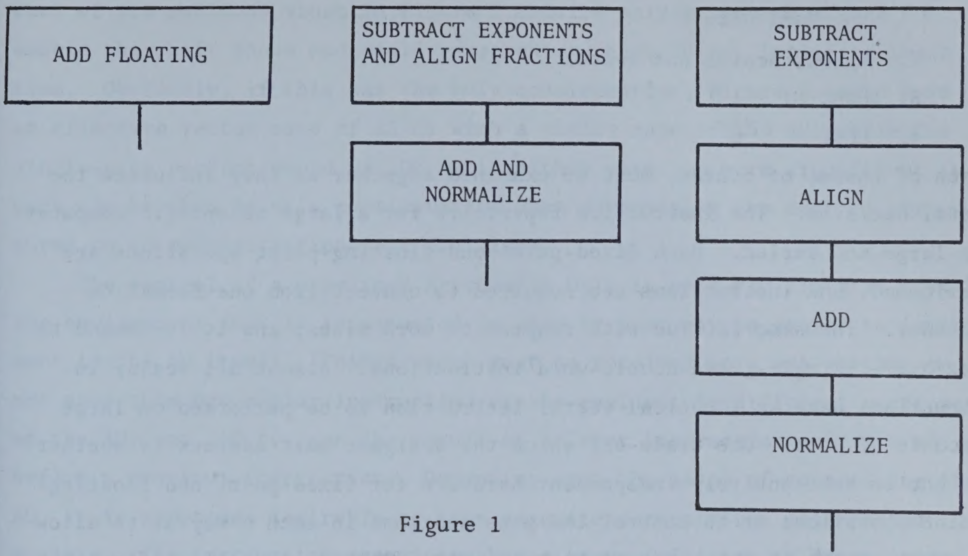


Figure 1

Other terms which need to be defined are Vector and Scalar instructions. A scalar instruction is one which operates on only one register operand and one memory operand to produce one result which normally would be placed in the addressed register. A vector instruction is one which operates upon a series of operands (usually two different areas of memory) and produces a series of results or, in special cases, a single result. Examples of this would be a Vector Add which would take *n* elements of Vector A, add those to corresponding *n* elements of Vector B, and produce a result Vector C which has *n* elements. A Vector Dot Product instruction, however, can multiply A_n by B_n and produce a single result which represents the sum of the individual products. When executing vector instructions, the basic measurement, or the effective vector rate, is stated in clocks per operand. The goal with a single pipeline is to operate upon a new set of operands each clock. For vectors which produce a result of *n* elements, an effective vector rate of one clock per element would require *n* clocks plus some time to fill the pipe. If an instruction uses the same pipe section for two clocks without allowing the next operand to enter, the vector rate is two clocks per operand.

DESIGN CONSIDERATIONS

Many factors influence the design of a pipelined Arithmetic Unit. Among these factors are:

1. Instruction Repertoire
    Fixed and Floating Point

Word Sizes

2. Speed, Scalar and Vector
3. Control
4. Cost

Each of these, of course, must be examined together as they influence the total decision. The instruction repertoire for a large scientific computer is large and varied. Both fixed-point and floating-point operations are performed, and instructions are required to convert from one format to another.. The same is true with respect to word sizes; and it is common to use half-, single-, and double-word instructions. Almost all scalar instructions have an identical vector instruction to be performed on large amounts of data. One trade-off which the designer must address is whether or not to have entirely independent hardware for fixed-point and floating-point operations or to control the pipe sections in such a way as to allow a great amount of the hardware to be common. One decision simplifies control, while the other lowers the cost.

A significant trade-off occurs when determining the proper pipeline structure for achieving the execution time of a scalar instruction opposed to a vector instruction. A simple example of this trade-off can be seen for a shift instruction. Assume that we want to shift a 64-bit word *n* places. A pipeline could be constructed which would perform the functions seen in Figure 2, or a single-pipe section could be designed to do the shift. The single section would require a decode of 64 and selection logic to select one of 64 inputs into each of 64 bits. This represents several levels of logic and might require in the neighborhood of 100 ns to perform.

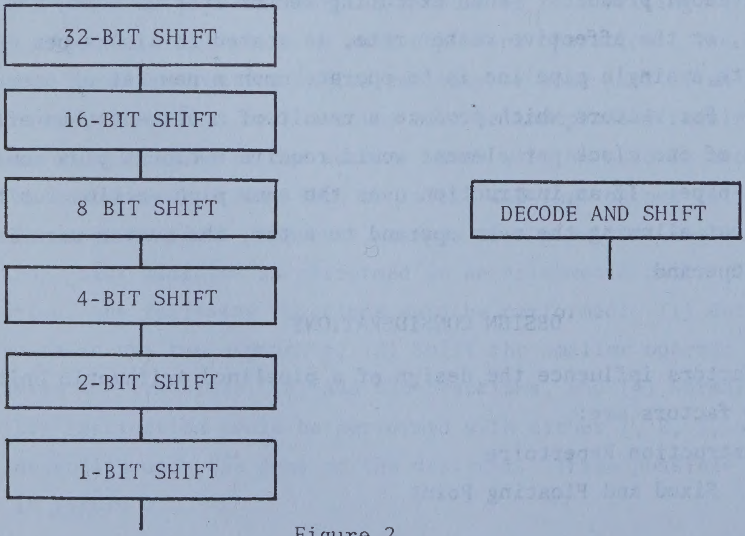


Figure 2

Each of the sections shown in Figure 2 require only a control signal to enable the shift shown and would represent perhaps 25 ns, including latch time. Obviously, if this was the only consideration, Figure 2 could have an effective vector rate of 25 ns with a scalar rate of 150 ns, while the single-pipe section would be 100 ns in either case. A more significant impact can be seen in this consideration when determining the desired structures to do multiplications and divisions.

The control of a pipelined Arithmetic Unit is probably dictated more by the environment that it is situated in than by what may be easier to implement in the AU itself. Things which must be resolved here are whether or not more than one scalar instruction may be resident in different sections of the AU; and, if so, can the result of a later instruction exit the AU before a previous instruction. Depending upon the means of controlling the AU, it is sometimes desirable to have separate controls for vectors and scalars. The instruction repertoire has a heavy influence in determining the way scalars can be overlapped in the pipeline since many instructions depend upon using some control output of a previous instruction.

Cost is always a significant consideration but is less of an issue in an Arithmetic Unit for a large scientific machine. There will be a large number of logic elements used to achieve the desired speed. Basic decisions concerning cost would include, though, the use of separate AU's for different functions such as fixed- and floating-point units. A major item to be included in cost, however, is the cost of maintaining a very large unit; and this influences heavily the choice of control philosophies. Also, to be considered here would be ease of modification to either expand present capabilities or to introduce new instructions.

THE ADVANCED SCIENTIFIC COMPUTER

To look at a specific design of a pipelined Arithmetic Unit, I will briefly describe the Texas Instruments Advanced Scientific Computer (ASC). Figure 3 shows that the Central Processor is composed of three units. The Instruction Processing Unit (IPU) fetches instructions, decodes the operation code, and develops the address of the memory operand. It also contains 64 addressable registers which may be selected to be operated upon. The functions of the IPU are pipelined. The Memory Buffer Unit (MBU) fetches operands from memory and stores appropriate results into memory. The MBU has complete control during vector instructions and calculates all memory addresses required for vectors. There are two levels of pipeline in the MBU. Also resident in the MBU is a Read-Only Memory (ROM), which is the basic control for the AU. The AU receives all operands from the MBU and provides results to either the IPU or MBU.

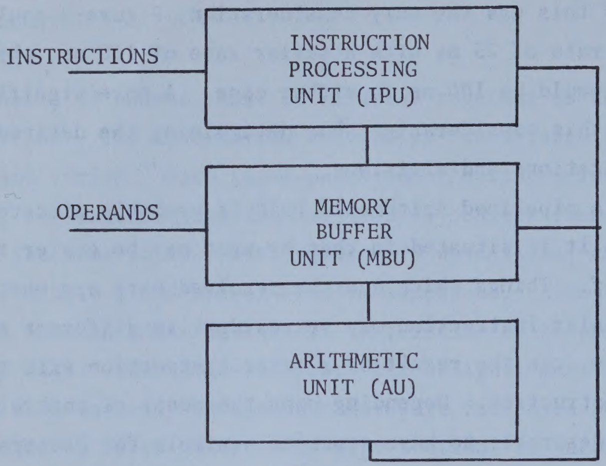


Figure 3

ARITHMETIC UNIT REQUIREMENTS

The basic requirements in designing the Arithmetic Unit for the ASC were:

1. Vectors performed at one clock per element.
2. Efficient scalar execution.
3. Very fast vector dot product.
4. Fixed- and floating-point formats.
5. Synchronous clock at best speed, approximately ten logic levels.
6. 64-, 32-, and 16-bit word sizes.

To satisfy requirements 1 and 2, it was determined to pipeline the Arithmetic Unit but to use as few pipe sections as practical to not unduly penalize scalar execution. The pipe sections could not contain too many logic levels, however, or the clock rate would not be able to produce desired execution times. Also, because of scalar execution requirements, it was decided to allow instructions to take varying paths through the AU and not restrict it to pass through each section to be operated upon or not. From a both a cost standpoint and ease of correction and expandability, a Read-Only Memory was selected as the basic control mechanism. The configuration shown in Figure 4 was the result.

This configuration was determined to have the fewest pipe sections required which would closely match the number of logic levels required to perform each function in the Central Processor. Following is a description of the general function of each section.

The Multiply section performs a full 32- by 32-bit multiplication. It produces output of two 64-bit numbers, referred to as a pseudosum and pseudocarry which, when added, will produce the desired product. Logic is included which allows iterative passes through the multiplier for double words and also to perform division by using an algorithm of successive multiplications.

The Accumulator receives inputs from the Multiply section and performs a parallel addition of the two 64-bit numbers. The Accumulator also has provision of feeding its output back through itself to perform double-length multiplications and divisions. This feedback path is also used to accomplish the Fixed-Point Vector Dot Product instruction.

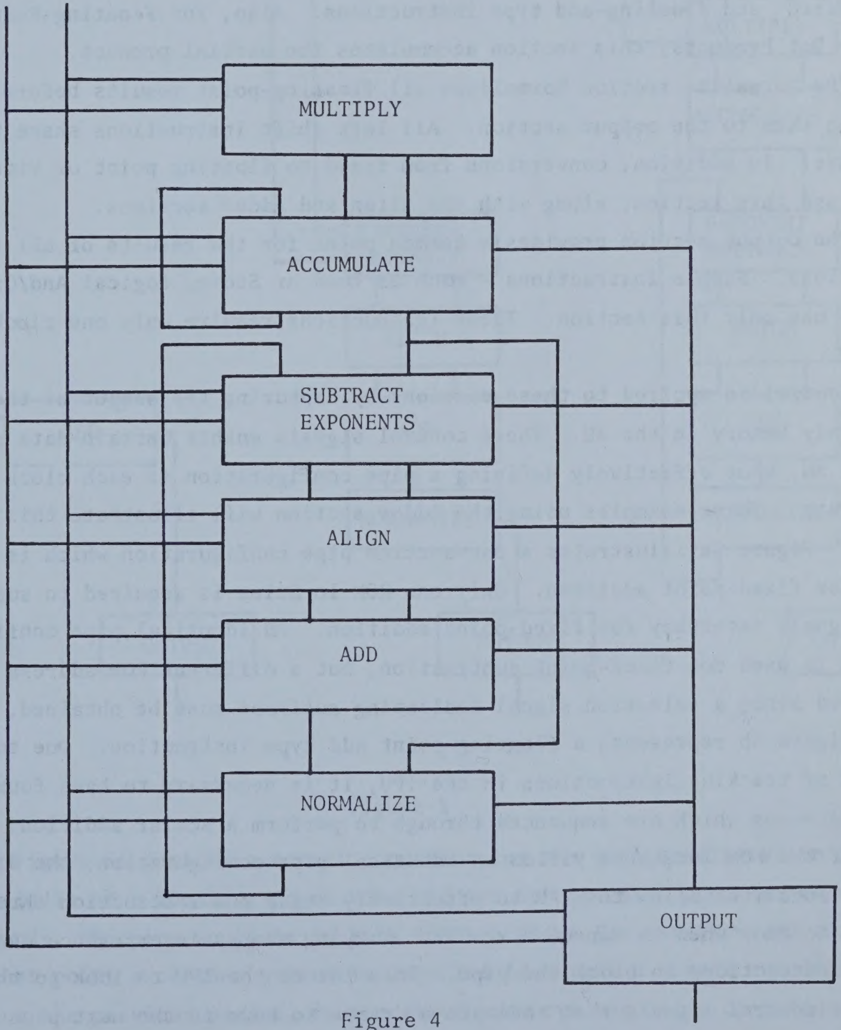


Figure 4

The Exponent Subtract section contains the logic required to determine the smallest number for floating-point operations - such as Add, Subtract, Add Magnitude, etc. - and to supply the amount of shift required to align the fractions. This logic is shared to do both Fixed- and Floating-Point Comparison instructions and, also, the special vector instructions such as Search for Largest and Search for Smallest.

The Align section performs the right shift required to align fractions for any floating-point instruction which uses the adder. The logic is shared to execute all right shifts - including logical, arithmetic, and circular shifts.

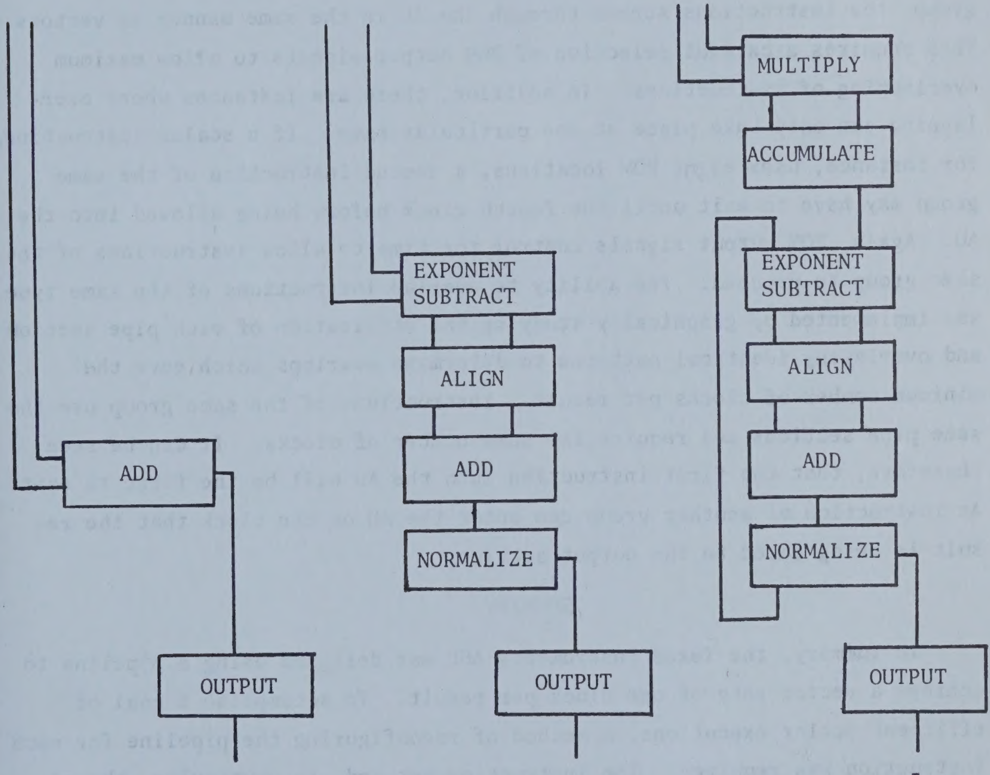
The Adder section consists of a double-level, look-ahead carry, propagating adder which can add two 64-bit inputs. The same hardware is used for both fixed- and floating-add type instructions. Also, for Floating-Point Vector Dot Products, this section accumulates the partial product.

The Normalize section normalizes all floating-point results before sending them to the output section. All left shift instructions share this hardware. In addition, conversions from fixed to floating point or vice versa use this section, along with the Align and Adder sections.

The Output section provides a common point for the results of all instructions. Simple instructions - such as Load or Store, Logical And/Or, etc. - use only this section. These instructions require only one clock in the AU.

Control is applied to these sections by capturing the output of the Read-Only Memory in the AU. These control signals enable certain data paths in the AU, thus effectively defining a pipe configuration at each clock if necessary. Three examples using the Adder section will illustrate this point. Figure 5a illustrates a one-section pipe configuration which is used for fixed-point addition. Only one ROM location is required to supply the signals necessary for fixed-point addition. An identical pipe configuration is used for fixed-point subtraction, but a different ROM address is accessed since a selection signal indicating *subtract* must be obtained.

Figure 5b represents a floating-point add type instruction. Due to the method of tracking instructions in the IPU, it is necessary to have four ROM addresses which are sequenced through to perform a scalar addition. Each of the ROM locations yields an identical pipe configuration; the different locations allow the IPU to effectively track the instruction through the AU to know when to expect a result. Varying memory interference can cause instructions to block the pipe. This forces the IPU to look to the ROM for control signals that indicate a desire to move to the next pipe section. By performing the appropriate complementation in the Exponent Subtract section, it is possible to present operands to the succeeding sections that can be operated upon independent of the actual instruction. Therefore, the last three ROM locations can be shared by a Subtract, Add Magnitude, Subtract Magnitude, etc. The ROM addresses for an Add may be 100, 101, 102, 103; while for a subtract, it could be 200, 101, 102, 103. The starting ROM address is supplied by the IPU directly from the instruction's operation code. Subsequent addresses are output from the ROM itself. This is an important aspect which I will discuss later concerning Same Group instructions.


ADD
OUTPUT
EXPONENT SUBTRACT
ALIGN
ADD
NORMALIZE
OUTPUT
MULTIPLY
ACCUMULATE
EXPONENT SUBTRACT
ALIGN
ADD
NORMALIZE
OUTPUT

5a 5b 5c

Figure 5

Figure 5c depicts a normal condition for the pipe configuration during the middle of a floating-point VDP. If the VDP operated upon ten thousand operands, the pipe would be in this configuration for ten thousand clocks. At that point a cleanup period is required to output a single result since, during the previous period, four partial products are being computed -

always one in each of the four sections of pipe that represent the floating-point addition. During this cleanup period, the pipe is configured several ways to allow proper data transfers.

Any time a vector instruction is resident in the AU, only that instruction is being operated upon. Vector instructions either use only one ROM control location or a short sequence of ROM code which is cycled through under control of the MBU. Part of the ROM output enables new operands into the AU and gates results out of the AU. Scalar instructions have a separate control consideration. Since all instructions use the same AU and there is a large number of instructions, only one type of instruction can be resident in the AU. However, if several instructions in sequence are of a common group, the instructions stream through the AU in the same manner as vectors. This requires a careful selection of ROM output signals to allow maximum overlapping of instructions. In addition, there are instances where overlapping can only take place at one particular time. If a scalar instruction, for instance, uses eight ROM locations, a second instruction of the same group may have to wait until the fourth clock before being allowed into the AU. Again, ROM output signals control the time to allow instructions of the same group to proceed. The ability to overlap instructions of the same type was implemented by graphically studying the utilization of each pipe section and overlaying identical patterns to determine overlaps which gave the minimum number of clocks per result. Instructions of the same group use the same pipe sections and require the same number of clocks. It can be seen, therefore, that the first instruction into the AU will be the first to exit. An instruction of another group can enter the AU on the clock that the result is being gated to the output section.

SUMMARY

In summary, the Texas Instruments ASC was designed using a pipeline to achieve a vector rate of one clock per result. To accomplish a goal of efficient scalar executions, a method of reconfiguring the pipeline for each instruction was required. The instruction set and, in particular, the requirement of a very fast Vector Dot Product produced a minimum requirement of seven pipe sections. Design considerations in the Instruction Processing Unit dictated a minimum of ten logic levels. It was determined that both fixed- and floating-point operations could be performed within that number of levels; and, thus, cost could be reduced by sharing logic. The selection of a ROM for the AU control has allowed new vector instructions to be added to the repertoire after the hardware was built and has provided a flexible control environment which can be easily documented and maintained.

MEMORY SYSTEMS FOR PARALLEL ARRAY PROCESSORS

DUNCAN H. LAWRIE
Department of Computer Science
University of Illinois
Urbana, Illinois 61801

ABSTRACT

This paper discusses the organization of memory systems for parallel array processors which allow conflict-free access to various slices of data and subsequent unscrambling of the data to place it in correct order for processing. The discussion centers on conflict-free data storage techniques and data alignment networks.

INTRODUCTION

By an array processor, we mean an array of N processors all of which execute a common instruction stream (e.g., ILLIAC IV). We also assume there are M memory modules and that these memories are cycled together to fetch or store M words. However, these memories need not all use the same memory address, i.e., each memory has its own independent address register.

One of the most significant factors in determining the performance of an array processor is the design of the memory system. It is not sufficient to merely provide a memory whose raw bandwidth is matched to the array of processors. We must also insure that data access normally occurs without memory conflict and that operand pairs can be brought together in the right processors. In this paper we will discuss these problems and present several possible solutions.

As an illustration of the problem, let us examine the case of a four processor system with four memories: N = M = 4. (Generally, we are only worried about larger systems, $N \geq 16$, but we will use N = 4 throughout in order to keep the examples manageable.) Assume we have an N x N matrix stored in the N memories as shown in Figure 1. (We shall only consider N x N matrices for the moment. Larger matrices may be partitioned into N x N pieces.) As we can see, access to rows of this matrix involves no memory conflict. However, all the elements of a column are stored in one memory so N memory cycles would be required to fetch or store a column.

Figure 2 illustrates another way of storing matrices. Here the matrix is skewed so that each row begins in the memory to the right of the row before it (1-skew). Now we can access either rows or columns without conflict. However, we can no longer access diagonals as we could using the scheme shown in Figure 1.

ACCESS REQUIREMENTS

In general, the access requirements of the memory system will depend on the type of problems to be run on the system. For example, solution of partial differential equations by implicit methods on rectangular coordinates generally requires only access to rows, while eigenvalue, eigenvector calculations require access to at least rows, columns, and diagonals. Figure 3 illustrates the access requirement with which we will be concerned in this paper. Access to rows, columns, and diagonals occur frequently in many application areas. Partial rows, columns, $\sqrt{N} \times \sqrt{N}$ square blocks, etc., are frequently used when large matrices are partitioned into $\sqrt{N} \times \sqrt{N}$ blocks rather than N x N blocks in order to improve the "resolution" of the processor array (c.f., Kuck [1]), as well as for other reasons.

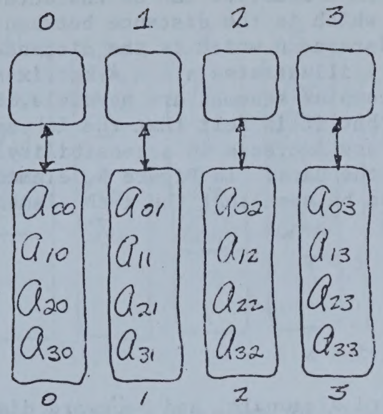

0
1
2
3
a_{00}
a_{10}
a_{20}
a_{30}
a_{01}
a_{11}
a_{21}
a_{31}
a_{02}
a_{12}
a_{22}
a_{32}
a_{03}
a_{13}
a_{23}
a_{33}
0
1
2
3

Figure 1. Four memories and processors: 0-skew storage

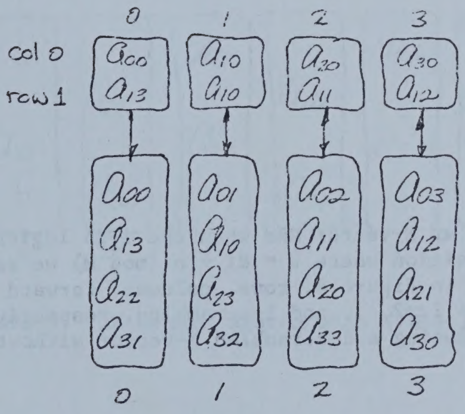

0
1
2
3
col 0
row 1
a_{00}
a_{13}
a_{10}
a_{10}
a_{20}
a_{11}
a_{30}
a_{12}
a_{00}
a_{13}
a_{22}
a_{31}
a_{01}
a_{10}
a_{23}
a_{32}
a_{02}
a_{11}
a_{20}
a_{33}
a_{03}
a_{12}
a_{21}
a_{30}
0
1
2
3

Figure 2. Four memories and processors: 1-skew storage

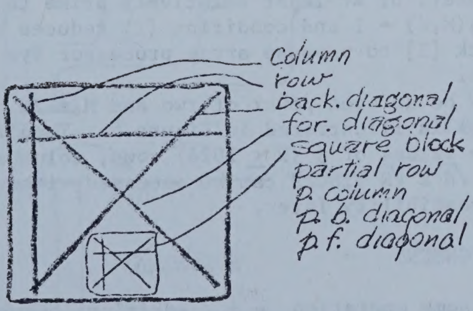

column
row
back. diagonal
for. diagonal
square block
partial row
p. column
p. b. diagonal
p. f. diagonal

Figure 3. Frequent access requirements in an N x N matrix

DATA STORAGE SCHEMES AND CONFLICT-FREE ACCESS

Schemes for storing N x N matrices in M memories can be characterized by two parameters: the *skew* distance α which is the distance between consecutive column elements and the *skip* distance β which is the distance between consecutive row elements. Figure 4 illustrates a 4 x 4 matrix stored in 5 memories with $\alpha = 2$, $\beta = 1$. More complex schemes are possible, for example where α and β are not constant, but it is felt that the increase in indexing difficulties are not offset by any increase in accessibility.

Now, consider fetching a column of the data. In Figure 4, elements of the second column are circled, and it can be seen that the i-th element of this column is located in memory μ where

$$\mu_{col} = 2i + 1 \pmod 5 \qquad 0 \leq i < N$$

In general, for arbitrary α and β,

$$\mu_{col} = \alpha i + c \pmod M \qquad 0 \leq i < N$$

More generally, for rows, columns, forward diagonals, and backward diagonals, the i-th element is stored in memory

$$\mu = di + c \pmod M \qquad 0 \leq i < N$$

where

$$d_{row} = \beta$$

$$d_{col} = \alpha$$

$$d_{f.d.} = \alpha + \beta$$

$$d_{b.d.} = \alpha - \beta$$

We call a vector of N elements an N-vector and when the i-th logical element is in the μ- th physical position where $\mu = di + c \pmod M$ we say the N-vector is d-scrambled. Thus, in Figure 4, rows, columns, forward diagonals and backward diagonals are 1, 2, 3, and 1-scrambled, respectively.

A sufficient condition for accessing a d-scrambled N-vector without conflict is that [2]

$$M \geq N \gcd(M,d) \tag{1}$$

where gcd(M,d) is the greatest common divisor of M and d. (We have not discussed $\sqrt{N} \times \sqrt{N}$ square blocks yet but notice that if $\alpha = \sqrt{N}$ and $\beta = 1$, then we can access these blocks as well.)

Notice that if M is a prime number, or at least relatively prime to $d = \alpha$, β, $\alpha + \beta$, and $\alpha - \beta$, then $\gcd(M,d) = 1$ and condition (1) reduces to $M \geq N$. This fact led Budnik and Kuck [3] to propose array processor systems with a prime number of memories.

Another possibility is to let N be an even power of two and M = 2N. Also, let $\alpha = \sqrt{N} + 1$ and $\beta = 2$. This is illustrated in Figure 5. Then it can be shown [2] that for reasonable values of N ($N \leq 1024$) rows, columns, forward and backward diagonals, and $\sqrt{N} \times \sqrt{N}$ blocks can be accessed without conflict. We shall return to this possibility later.

DATA ALIGNMENT

Now suppose we wish to perform some operation, e.g., addition, between two N-vectors. If we wish to add two columns, then we must *shift* one column relative to the other in order to align them. For example, in Figure 4, we must shift column 1 one position to the left in order to add it to

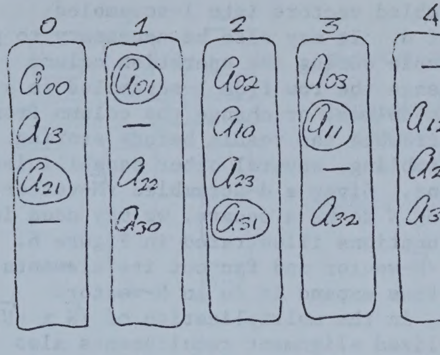

0
1
2
3
4
a_{00}
a_{13}
a_{21}
—
a_{01}
—
a_{22}
a_{30}
a_{02}
a_{10}
a_{23}
a_{31}
a_{03}
a_{11}
—
a_{32}
—
a_{12}
a_{20}
a_{33}

Figure 4. A 4 x 4 stored in five memories

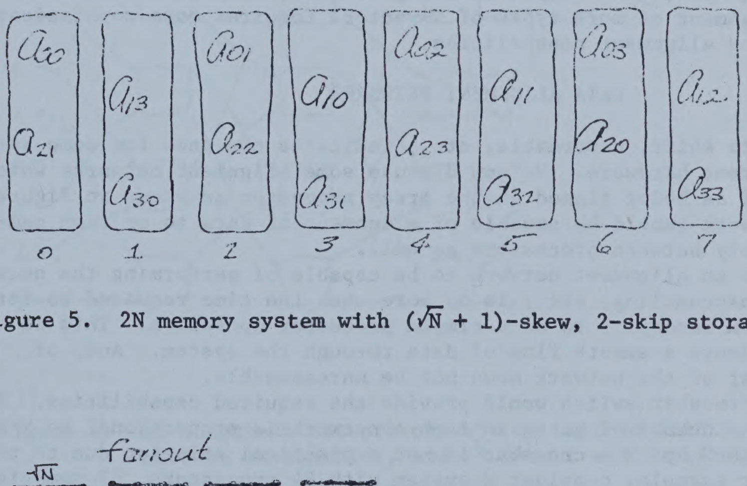

a_{00}
a_{21}
a_{13}
a_{30}
a_{01}
a_{22}
a_{10}
a_{31}
a_{02}
a_{23}
a_{11}
a_{32}
a_{03}
a_{20}
a_{12}
a_{33}
0
1
2
3
4
5
6
7

Figure 5. 2N memory system with ($\sqrt{N}$ + 1)-skew, 2-skip storage

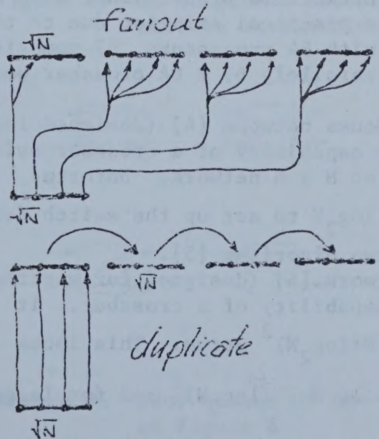

fanout
$\sqrt{N}$
$\sqrt{N}$
$\sqrt{N}$
duplicate
$\sqrt{N}$

Figure 6. Patterns expanding $\sqrt{N}$-vectors into N-vectors

column 0. However, to add a row to a column we find that a row is 1-scrambled while a column is 2-scrambled. Thus, in addition to a shift, we have to change a 2-scrambled vector into 1-scrambled vector. Thus, in general, we must be able to change d-scrambled vectors into 1-scrambled vectors, at least for certain values of d. It may also be necessary to perform the inverse operation, as for example during the operation column = row + column. Here we can either change the row from 1-scrambled to α-scrambled, add, and store directly to a column; or change the column from α-scrambled to 1-scrambled, add, and scramble the result before storing.

In addition to shifting and unscrambling, several other capabilities are also important for some applications. Given a d-scrambled $\sqrt{N}$-vector (partial row, column, etc.) and assuming N to be a square, we may need in addition to unscrambling, one of the functions illustrated in Figure 6. That is, we may need to unscramble an $\sqrt{N}$-vector and fan out its elements $\sqrt{N}$ times or duplicate it $\sqrt{N}$ times and thus expand it to an N-vector. (These requirements occur, for example, in the multiplication of $\sqrt{N} \times \sqrt{N}$ partitions in $\sqrt{N}$ steps.) Other specialized alignment requirements also exist, such as the shuffle connection for the fast Fourier transform.

The particular alignment requirements of an array processor can only be determined by a study of the applications to be handled by the processor. (Or the applications may be determined by the capabilities of the processor.) For example, if access and alignment of rows and columns is all that is required, then simple shifting alignment may be sufficient. Access and alignment of more types of N-vectors requires more sophisticated data storage and alignment capabilities.

DATA ALIGNMENT NETWORKS

The need to shift, unscramble, etc., indicates the need for some specialized alignment hardware. We now discuss some alignment networks which we can think of as being placed in the array processor as shown in Figure 7. Such a network should be capable of alignment of data to or from memory, and possibly between processors as well.

We require an alignment network to be capable of performing the necessary shifts, unscrambling, etc., in no more than the time required to fetch an N-vector from memory or to do a simple processor operation. This is necessary to insure a smooth flow of data through the system. And, of course, the cost of the network must not be unreasonable.

An M x N crossbar switch would provide the required capabilities. Unfortunately, the number of gates in such a network is proportional to $M \cdot N$ and so for large N and M a crossbar is not a practical solution due to the gate cost. For example, consider a system with 64 processors, 67 memories, and data words of 32 bits. A 32-bit, word parallel, 67 x 64 crossbar would require an order of 137K gates.

Another possibibility is the Clos or Benes network [4] (designed for telephone switching). This network has the capability of a crossbar switch but has only an order of $N \log_2 N$ gates for an N x N network. Unfortunately, it requires time on the order of $N \log_2 N$ to set up the switch for a given alignment function using the best known algorithm [5].

A third possibility is the Batcher network [6] (designed for sorting applications.) This network also has the capability of a crossbar. It works in order $(\log_2 N)^2$ time and has order $N(\log_2 N)^2$ gates. This looks more appealing but unfortunately, for N small, $N \simeq (\log_2 N)^2$ and for larger N the number of gates is again prohibitive.

Another possibility which has recently been identified is shown in

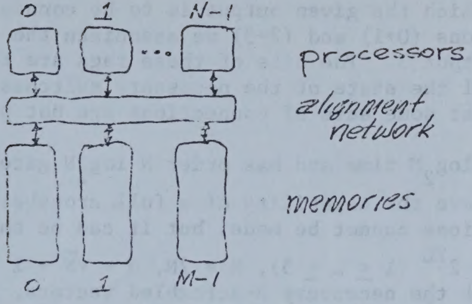

0
1
...
N-1
processors
alignment network
memories
0
1
M-1

Figure 7. An array processor with alignment network

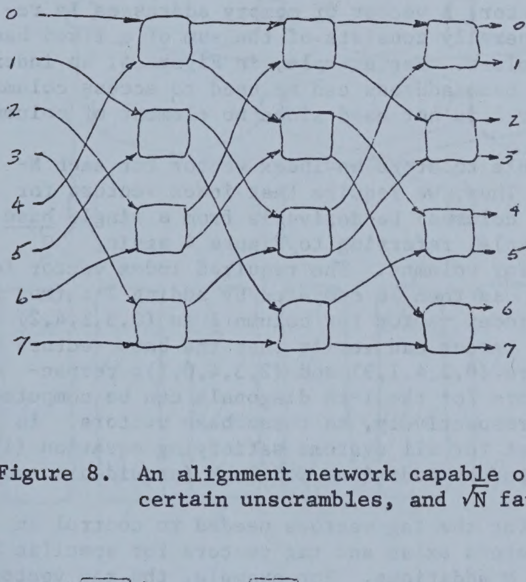

0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
7

Figure 8. An alignment network capable of uniform shifts, certain unscrambles, and $\sqrt{N}$ fanout and duplication

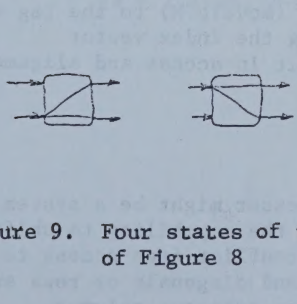


Figure 9. Four states of the switching elements of Figure 8

Figure 8. Each box in Figure 8 has the capabilities for switching its inputs to its outputs as shown in Figure 9. Control of this network is straightforward. With each output port we associate a binary tag which is the number of the input port to which the given output is to be connected. For example, to make the connections (0→1) and (2→3) we associate the tags 000 with output 1 and 010 with output 3. The bits of these tags are then used in a direct manner to control the state of the necessary switches in the network. Notice, however, that some sets of connections are not possible, e.g., (0→1) and (4→0).

This network works in order $\log_2 N$ time and has order $N \log_2 N$ gates. As mentioned above, it does not have the capability of a full crossbar switch since some sets of connections cannot be made, but it can be shown [2] that given a system where $N = 2^{2L}$ $(1 \leq L \leq 5)$, $M = 2N$, $\alpha = \sqrt{N} + 1$ and $\beta = 2$, this network can unscramble the necessary d-scrambled vectors, do fan outs or duplicates, and do any uniform shift, all at the same time. Such a system is illustrated in Figure 10 for $N = 4$. The capabilities and construction of this network are discussed in more detail in [2].

INDEXING

In order to access an N-vector, a vector of memory addresses is required. This address vector generally consists of the sum of a fixed base address and a vector of index values. For example, in Figure 4, an index vector of (0,3,1,4,2) plus some base address can be used to access column 0. (The index value 4 in memory 3 is not used since no element of column 0 is in this memory.)

It is probably not reasonable to store an index vector for each N-vector that we wish to access. Thus, we require that index vectors for each N-vector in a class (e.g., columns) be derivable from a single base vector for that class. For example, referring to Figure 4 again, (0,3,1,4,2) is the base vector for columns. The required index vector for the i-th column $(0 \leq i < N - 1)$ can then be computed by adding $2 \cdot i \pmod 5$ to the base vector. Thus, the index vector for column 1 is (0,3,1,4,2) + 2 (mod 5) = (2,0,3,1,4). The reader can verify that the base vectors for forward and reverse diagonals are (0,2,4,1,3) and (2,3,4,0,1), respectively, and that the index vectors for the i-th diagonals can be computed by adding (modulo 5) 3i and i, respectively, to these base vectors. In general, these base vectors exist for all systems satisfying equation (1), and index vectors can be generated by additions modulo M (provided α and β are constant).

A similar situation holds for the tag vectors needed to control an alignment network. Base tag vectors exist and tag vectors for specific N-vectors are generated by modulo M additions. For example, the tag vector for column 0 in Figure 4 is (0,2,4,1). This would cause data from memory 0 to go to processor 0, memory 2 to processor 1, 4→2, and 1→3. The tag vector for column 1 is then formed by adding 1 (modulo M) to the tag vector for column 0. The reader can verify that using the index vector (2,0,3,1,4) and tag vector (1,3,0,2) will result in access and alignment of column 1 $(a_{01}, a_{11}, a_{21}, a_{31})$.

SUMMARY

A minimal configuration for an array processor might be a system with the same number of memories as processors, and the capability to shift data one position right or left. This would allow conflict-free access to rows and columns (or, via storage remappings, rows and diagonals or rows and square blocks) and alignment of adjacent rows or adjacent columns.

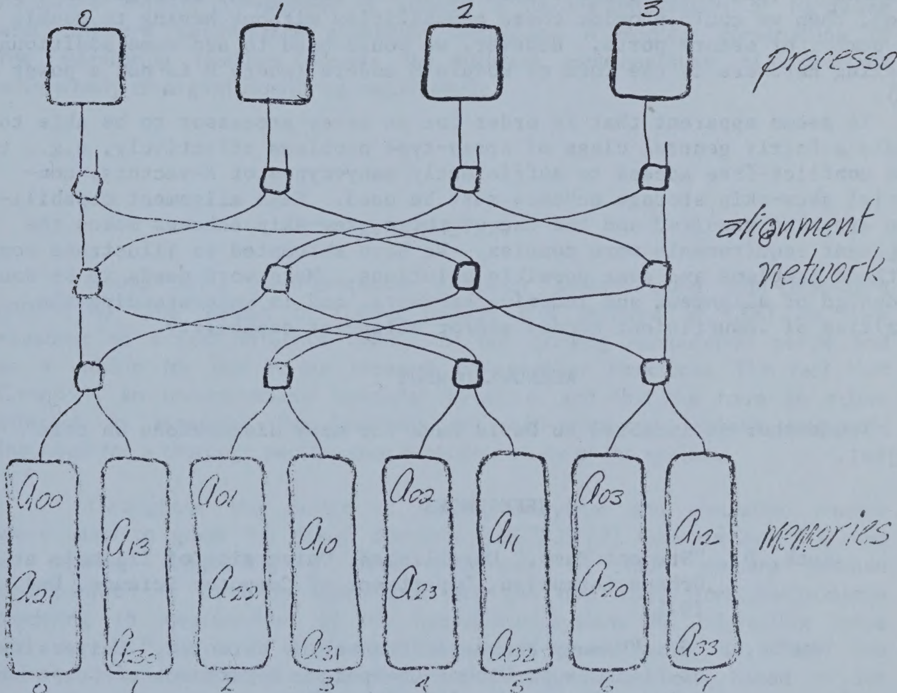

0
1
2
3
processors
alignment network
a_{00}
a_{21}
a_{13}
a_{30}
a_{01}
a_{22}
a_{10}
a_{31}
a_{02}
a_{23}
a_{11}
a_{32}
a_{03}
a_{20}
a_{12}
a_{33}
memories
0
1
2
3
4
5
6
7

Figure 10. N processor, 2N memory system with alignment network and $\sqrt{N}$ + 1 skew, 2-skip storage

Alignment of non-adjacent rows or columns, or of rows with columns, would require repetitive shifts and performance would degrade accordingly.

A more desirable configuration would include a device which could perform an arbitrary distance uniform shift in fixed time (e.g., [7]). This would allow alignment of arbitrary rows and/or columns. One can make arguments, however, that such a device must contain at least an order of N log N gates and require an order of log N time. For this same price in gates and time we can build a network (Figure 8) which can simultaneously unscramble certain d-scrambled vectors, fanout/duplicate, and do arbitrary uniform shifts. And by using this network and 2N memories where N is an even power of two, we can provide a system which allows access and alignment of rows, columns, diagonals and square blocks, and fanout/duplication of $\sqrt{N}$-vectors.

If we could design a similar network which worked with a prime number of memories (and there doesn't seem to be any reason why this couldn't be done), then we could provide these capabilities without having to double the number of memory ports. However, we would need to add some additional indexing hardware in the form of modulo M adders (where M is not a power of two).

It seems apparent that in order for an array processor to be able to handle a fairly general class of array-type problems effectively, e.g., to have conflict-free access to sufficiently many types of N-vectors, non-trivial skew-skip storage schemes must be used. Data alignment capabilities are also required and the use of these skew-skip schemes makes the alignment requirements more complex. We have attempted to illustrate some of these problems and some possible solutions. More work needs to be done in design of alignment and indexing hardware, and in understanding the penalties of insufficient access and/or alignment capability.

ACKNOWLEDGMENT

The author is indebted to David Kuck for many discussions on this subject.

REFERENCES

[1] Kuck, D., "Student Memo," Unpublished, University of Illinois at Urbana-Champaign, Department of Computer Science, Dec. 1971.

[2] Lawrie, D. H., "Memory-Processor Connection Networks," University of Illinois at Urbana-Champaign, Department of Computer Science Report No. 557, Feb. 1973.

[3] Budnik, P. and D. Kuck, "The Organization and Use of Parallel Memories," IEEE Transactions on Computers, Vol. C-20, Dec. 1971, pp. 1566-1569.

[4] Benes, V. E., Mathematical Theory of Connecting Networks and Telephone Traffic, Academic Press, N. Y., 1965.

[5] Opferman, D. C. and N. T. Tsao-Wu, "On a Class of Rearrangeable Switching Networks," Bell System Technical Journal, Vol. 50, May-June 1971, pp. 1579-1618.

[6] Batcher, K. E., "Sorting Networks and Their Applications," Proceedings of the Spring Joint Computer Conference, 1968, pp. 307-314.

[7] Semmelhaack, C. F., "Method and Apparatus for Routing Data Among Processing Elements of an Array Computer," U. S. Patent No. 3,582,899, June 1, 1971.

THE PERFORMANCE MONITOR FOR C.mmp (A MULTI-MINI-PROCESSOR)

S. H. FULLER and D. K. STEVENSON ⊗
Carnegie-Mellon University
Pittsburgh, Pennsylvania

ABSTRACT

The performance monitor for C.mmp, a multi-processor computer system, is a Simula-like programming facility that includes a programmable hardware monitor component to provide a flexible integration of hardware and software performance evaluation techniques. A modular construction of the monitoring facilities permits the efficient implementation of processes with widely divergent monitoring requirements.

INTRODUCTION

C.mmp is a multi-processor computer system using mini-computers (PDP-11's). The design of C.mmp [WB72] was initiated at CMU for two reasons: as a cost effective facility for our growing computation needs and as a vehicle for use in our research on computer structures. The fact that C.mmp is an unconventional computer structure, and that we have an active interest in assessing the feasibility of multi-processor systems, suggests the need for a thorough performance evaluation study of the system.

Throughout the design of C.mmp, analytical and simulation models were used to guide the design decisions [McC73,BF73]. Construction of C.mmp is nearing completion; a three processor by four memory module configuration is currently operational. We can now shift from performance modeling to measurement of the operational system. An interesting issue will be to return to the analytical models used earlier to check the validity of their assumptions and to see if other assumptions, based on our measurements, can provide more accurate models.

This paper begins with a brief overview of the organization of C.mmp and the basic principles of HYDRA, the operating system that is being developed for C.mmp [Wu72]. While reading these sections, note the distinctly different character a multi-processor like C.mmp has in contrast to the "vector" processors such as ILLIAC IV and the CDC STAR. It will be important to compare a multi-processor like C.mmp to these "vector" processors as well as to the more conventional uni-processors. The hardware monitor, treated as a sub-system of the performance monitor, is presented in some detail. The paper ends with an example of how a user might invoke a variety of the capabilities of the performance monitor.

⊗ This work was sponsored by the Advanced Research Projects Agency of the Office of the Secretary of Defense (F44620-73-0074) and is monitored by the Air Force Office of Scientific Research.

C.mmp: AN OVERVIEW

The hardware configuration of C.mmp is that of the archetypal multi-processor; see Figure one. A cross-point switch, Smp, allows a direct connection between any processor and any shared memory module, Mp, on a per-reference basis, and resolves conflicts between processors that request the same memory module. A second switch, Skp, allows connection between the processors (and hence the shared memory) and shared device controllers. Communication among processors is in the form of interrupts at any of several priority levels.

Each processor system is a complete computer with its own primary memory, controllers and devices. The local memory and devices serve both to reduce bandwidth requirements and to allow completely independent operation. The specific components shown in Figure one are explained in more detail below:

K.clock: A central clock is readable by all processors and broadcasts a time base for local interval timing.

K.interrupt: Any processor may generate an interrupt to any subset of the Pc configuration.

Smp: This switch handles transfers between the Pc and Mp configurations.

Mp: The primary memory consists of (up to) 16 modules of (up to) 65k, 16 bit, words.

Pc: The central processing elements are slightly modified versions of the DEC PDP-11.

Dmap: Dmap maps the addresses generated by the processor and device controllers into addresses in the shared primary memory.

A NOTE ON THE OPERATING SYSTEM

Three aspects of the design of the operating system for C.mmp have guided the design of the monitoring system. To some extent this is due to the monitoring system being a (compatible) resource of the system. Also the principles involved in the design of an operating system are just as applicable to the design of a monitoring facility. The relevant considerations are listed below:

1. IMPLEMENTATION METHODOLOGY: The implementation methodology for the operating system is a cross between structured programming as advocated by Dijkstra and others [DDH72] and the modularization philosophy of Parnas [Pa72]. An immediate implication of this is that the location and format of information used by the monitoring facility may not be known to the monitor, so information must be acquired from the relevant module via a

---

⊗ We use the PMS notation [BN71] throughout this paper. It is based on seven primitive component types: P-processor, M-memory, S-switch, L-link,K-controler,T-transducer, and D-data operator. A computer structure built from primitive PMS components is represented by C; C.mmp denotes a multi-mini-processor Computer.

procedure call. Thus the alteration of any component has no effect on the monitor itself, since its interface with the monitor is maintained by the component. This is especially desirable in an evolving, experimental environment.

2. REJECTION OF STRICT HIERARCHICAL LAYERING: The notion of a strict hierarchically layered system has become popular since first described by Dijkstra for the THE system [Dij68]. While the system as viewed by any single user is to be hierarchically structured, this notion is rejected by the operating system as a global design criterion, since such an approach would severely limit the flexibility available to the high-level user and would strangle experimentation. Hence the design of the monitoring facility cannot assume any such layering of the systems for whch it may be used to monitor.

3. PROTECTION: Flexibility and protection are closely related, but are not inversely proportional. It is essential for protection to exist in a uniform manner throughout the system. It should not be applied only to specific entities (e.g. files) and not to others (e.g. resources when accessed by a monitor). The protection scheme is based on the idea of capabilities (in the sense of Dennis [DvH66]).

These goals, among others, dictate that at the heart of the operating system is a collection of facilities of "universal applicability," a set of mechanisms from which an arbitrary set of operating system facilities and policies can be conveniently constructed. Moreover, lest the flexibility be constrained at any instant, it should be possible for an arbitrary number of systems created from these facilities to co-exist simultaneously. The collection of such facilities is called the kernel.

Clearly, some of the monitoring functions must be supplied by the kernel. It is one of the major characteristics of the monitoring system for C.mmp that not all of the monitoring capabilities reside within the kernel.

THE MONITORING FACILITY

A monitoring system for C.mmp can be conceptually divided into three parts: the hardware monitor sub-system, the kernel-resident monitor, and a non-kernel "deus ex machina." The third part must be created for each monitoring process; to aid in this construction, a high-level language is supplied for expressing how the hardware monitor and kernel-resident portion of the monitoring facility are to be used.

The C.mmp Hardware Monitor

The fact that C.mmp is a multi-processor has a significant impact on the design of a hardware monitor [FSW73]. First, a multi-processor aggravates certain problems, adding to the measurements required to monitor the system adequately. Second, a multi-processor eliminates the constraint that a software monitor cannot significantly utilize a central processor without perturbing the system it is monitoring. This reduces the hardware versus software monitor controversy to a series of engineering trade-offs.

The C.mmp hardware monitor is a peripheral device connected to a PDP-11 Unibus, similar to a teletype or disc unit. The processor on this Unibus, referred to below as the supervisory processor, P.sup, is the only processor which can address the registers in the hardware monitor directly. The monitor is connected to a second Unibus and it is this processor, P.host, which is monitored. All communication between the processor, peripherals and memory is via the Unibus, conveniently concentrating most of the signals of interest to the hardware monitor. A few extra signals, generated internally by the PDP-11, are also accessible to the hardware monitor. Provision is also made for a number of general purpose, high-impedance probes.

The principal sections of the hardware monitor are shown in Figure two. The primitive event detector senses events at a Unibus cycle (i.e. memory fetch) level. It is composed of two types of modules: comparitors and bit masks. A comparitor has an internal register (set by P.sup) and produces two output signals, "input = register" and "input < register." A bit mask has two internal registers (set by P.sup): one specifies care / don't care conditions for each bit, the other specifies the expected bit pattern. A single output line indicates if the input satisfies the specified mask. Signal inputs for the bit mask and comparitor modules are arranged in four groups: Unibus address, Unibus data, control group and miscellaneous. The control group includes some Unibus control signals, the general purpose, high-impedance probes, and signals used for measurement of congestion in the cross-point switch. In the current implementation of the monitor, four primitive events may be defined simultaneously.

An accumulated event is the nth occurrence of an associated primitive event. The event accumulator modules within the accumulated event detector are composed of a register set by P.sup and a synchronous counter. The counter is incremented whenever the corresponding primitive event is detected. When the counter and the register contents are equal (which defines an accumulated event) all dynamic information is saved. The control register specifies new states for the internal boolean flags and may specify that P.host or P.sup be interrupted.

The event recorder buffers the data saved when an accumulated event is detected and ultimately transfers the required data to main storage. All the external signals available for primitive event detection, the value of real time and the internal counts of primitive events are saved in a double buffered MOS shift register store. Corresponding to each accumulated event is a register which specifies the data words to be transferred to main storage via the P.sup Unibus.

The interface to P.host is principally passive, providing data and timing signals for primitive event detection. However the interface can cause an interrupt to P.host. The interface to P.sup is used to set the internal registers of the hardware monitor and to transfer data to main memory.

It is intended that eventually several monitors will be provided for the measurement of C.mmp. The clocks within each monitor will be synchronised to facilitate comparison of measurements.

The Kernel-Resident Monitoring System

The kernel-resident portion of the monitor consists of two parts. The first is distributed throughout the kernel; it writes monitoring information about the kernel's operation into a special core-resident table. Presumably this information (which is continually being gathered) would be of continuing interest to the kernel in realizing some form of dynamic policy selection. Typical of such information currently envisioned being gathered by this part of the kernel-resident monitor would be: the length of various (critical) kernel system queues, timing and counts from the I/O devices, etc.

The second conceptual part of the kernel-resident monitor is a set of procedures which when invoked return more exotic information. For example: given a list of process names a procedure that returns the ones that are currently on the list of feasible processes to run.

Also in this second part of the kernel-resident monitor would be the set of self-monitoring routines supplied by the sub-systems themselves which report their status through the kernel. Each software facility would also provide an appropriate set of monitoring procedures. For example, a file management sub-system might supply procedures that would return how many files were accessed since its last call, what type of accesses were made, where the files were located, etc. All sub-system monitoring procedures may be called through the kernel so that any one process can potentially monitor the entire system.

The Deus ex Machina

A third part of the monitor does not reside in the kernel; it is tailored to fit the particular monitoring objective and is easily changed. It, too, consists of two conceptual parts (conceivably separate processes): an orderly sequence of procedure calls on the kernel to generate information and a collector / data-reduction component to process the information. Note that only the first part need be core-resident during the actual monitoring of the system.

Of course, with this division of the monitoring system into a kernel and non-kernel part, there is no reason why there cannot be several distinct monitors operating on C.mmp simultaneously.

The strong affinity between a monitor and a simulation suggests that the sequence of procedure calls could most naturally be couched in a simulation language. A high-level implementation language for the PDP-11 (called Bliss-11) was selected as the base language for describing the non-kernel part of a monitor. One of the inviting characteristics of this language is a powerful macro facility. The high-level monitoring language is implemented as macro definitions. Not only does this allow us to use an existing (optimizing) compiler, but it also makes the monitoring language "infinitely extensible." SIMULA-like constructs [DN66] are available via macro definitions. In addition, further macros are included that are specially designed for describing "primitive events" useful for monitoring purposes. These camouflage the procedure calls and make the high level monitoring language reasonably self-documenting. The example in the next section illustrates this characteristic; any obscurity can be traced to the arbitrary I/O structure.

One should keep in mind that a monitoring program (even one written in a simulation language) will be considerably simpler than the usual simulation; hopefully the overhead of the simulation mechanisms will in such a case be worth the benefits of clarity and ease of use.

The collector / data-compressor component of a monitoring system for C.mmp can be tailored for each construction of a monitor. These could vary from real-time selection for display purposes at the operator's console to rote depositing onto magnetic tape for later processing.

EXAMPLE

The publication conventions are as follows: Boldface (e.g. **Begin**) denotes reserved words in Bliss-11 and hence in the higher-level monitoring language. Capitalization (e.g. Timeoday) denotes macros supplied by the monitoring language.

The name of a variable denotes its address. A dot before the name denotes the contents of the address specified. In this example the user has defined two macros for himself: Record which sends a message to the data collector process; its first argument is the starting address of the message, its second argument is the length of the message. The other macro, Code, indicates the type of message being sent.

```
Monitor

On (KC17) %PL/1-like ON condition executed when Kernel Call #17 is made%
          Begin
          Local arry[3];
          arry[0]← Code(7);
          arry[1]←.Timeoday; % real time of day %
          arry[2]←User; % process identification %
          Record(arry,3)
          End
No; % end of On condition %

Routine A=Begin
          Own arry[201],i,t;
          While .Time Leq .Maxtime Do
                  Begin
                  Incr i From 1 To 199 By 2 Do
                          Begin
                          arry[.i]←Qlength(Disk0); % length of queue
                                  for disk 0 %
                          arry[.i+1]←.Timeoday;
                          Hold(Uniform(5.0 Milsec,10.0 Milsec,t))
                          End;
                  arry[0]←Code(1);
```

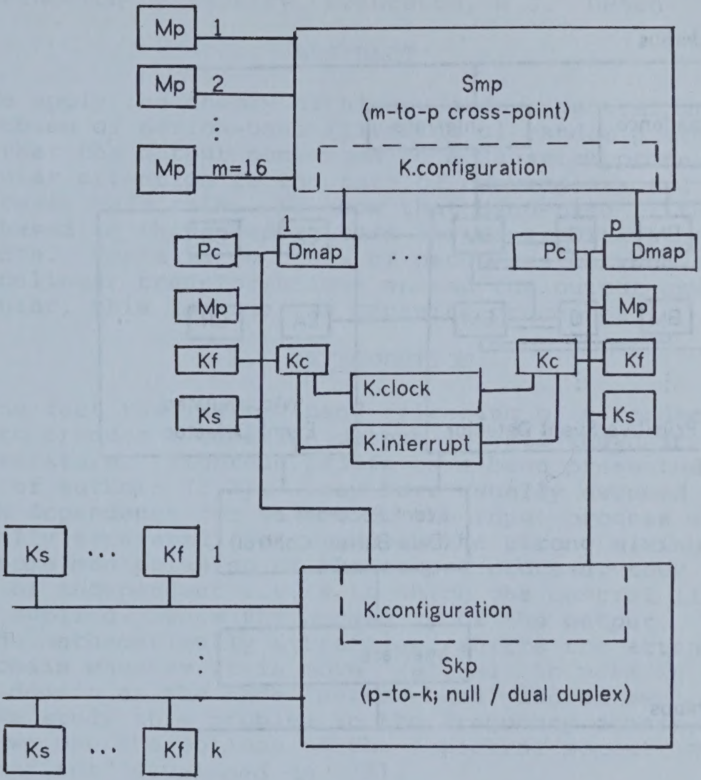
```
                Record(arry,200);
                Hold(5.0 Sec) % wait 5 seconds %
        End;

Routine B=Begin
        Own arry[4];
        arry[0]←Code(2);
        While .Time Leq 0.2 Min Do
                Begin
                arry[1]←Qlength(Disk1);
                arry[2]←.Timeoday;
                arry[3]←Feasiblelist(Lngth); % Feasiblelist is list of
                                            processes that can run %
                Record(arry,3);
                Hold(300.0 Milsec)
                End
        End;

Routine C= % manages the enabling of On condition during monitoring %
        While .Time Leq .Maxtime Do
                Begin
                Own s,t;
                Enable (KC17);
                Hold (Uniform (50.0 Milsec,100.0 Milsec, t));
                Disable (KC17);
                Hold (2.5 Sec);
                s←Code(5);
                Record(s,1)
                End;

Maxtime←0.5 Min; % maximum length of time monitoring program will run %
Parallel (A(),B(),C()) % A,B,C run in parallel %
Rotinom; % end of monitor module %
```

CONCLUSION

The multiple processor aspect of C.mmp allows great flexibility in balancing a hardware-software monitoring system. The monitoring facility for C.mmp exploits this freedom in a modular design which, for the user, appears as a high-level Simula-like monitoring language. Because the atmosphere surrounding C.mmp is experimental, the design of the monitoring system has had to accomodate the fact that the operating systems it is to monitor are evolving; indeed, the information it collects will be used to direct this evolution. This dictates that the performance monitor be capable of radical tailoring to meet future needs not yet determined (or expected). It also suggests that the interface between the performance monitor and the sub-systems be rigidly preserved, no matter how the sub-systems may change. Any unbearable overhead in monitoring frequent "primitive" events by a software approach will be obviated by the software programmable hardware monitor.

BIBLIOGRAPHY

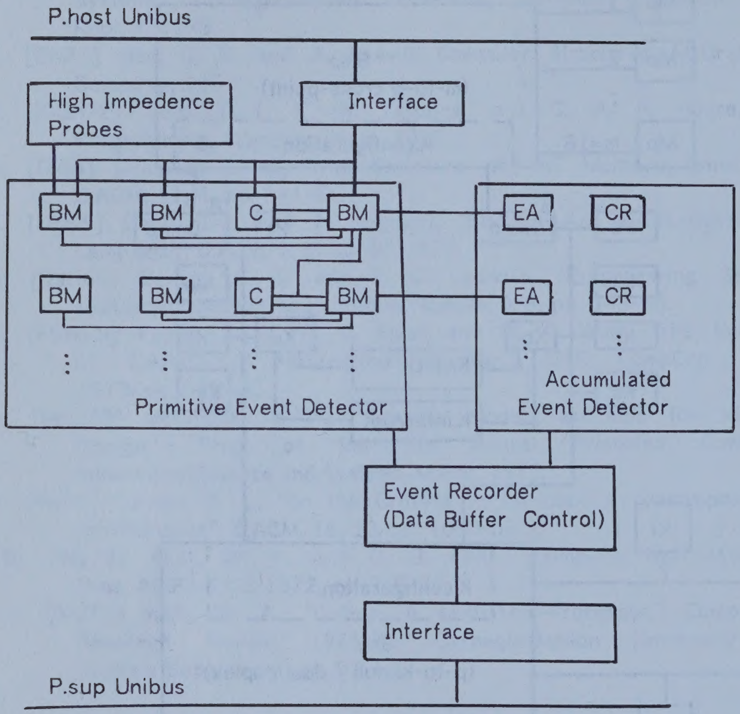
1. [BF73] Bhandarkar, D. P. and S. H. Fuller, **A Survey of Techniques for Analyzing Memory Interferance in Multi-Processor Computer Systems**, Carnegie-Mellon University, Computer Science Department Report, 1973.
2. [BN71] Bell, C. G. and A. Newell, **Computer Structures**, McGraw Hill Book Company, 1971.
3. [DDH72] Dahl, O.-J., E. W. Dijkstra, and C. A. R. Hoare, **Structured Programming**, Academic Press, 1972.
4. [Di68] Dijkstra, E. W., "The Structure of THE Multiprogramming System," **C.ACM, 11**, 5, pp. 341-55.
5. [DN66] Dahl, O.-J. and K. Nygaard, "SIMULA--An ALGOL-Based Simulation Language," **C.ACM**, 9, 9, pp. 671-678.
6. [DvH66] Dennis, J. B. and E. C. vanHorn, "Programming Semantics for Multiprogrammed Computations," **C.ACM**, 9, 3, pp. 143-55.
7. [FSW73] Fuller, S. H., R. A. Swan and W. A. Wulf, "The Instrumentation of C.mmp, A Multi-(Mini)-Processor," **IEEE CompCon '73**, March, 1973, pp. 173-76.
8. [McC73] McCredie, J. W., "Analytic Models as Aids for Multiprocessor Design," **Proc. of the 7th Annual Princeton Conference on Information Science and Systems**, March, 1973.
9. [Pa72] Parnas, D. L., "On the Criteria to be Used in Decomposing Systems into Modules", **C.ACM**, 15, 12, pp. 1053-58.
10. [WB72] Wulf, W. A. and C. G. Bell, "C.mmp--A Multi-Mini-Processor," **Proc. AFIPS FJCC 1972**, pp. 765-77.
11. [Wu72] Wulf, W. A., "C.mmp: A Multi-Mini-Processor," **Computer Science Research Review 1971-72**, Carnegie-Mellon University, Computer Science Department, pp. 37-69.


Mp 1
Mp 2
Mp m=16
Smp
(m-to-p cross-point)
K.configuration
1
p
Pc
Dmap
Pc
Dmap
Mp
Mp
Kf
Kc
Kc
Kf
K.clock
Ks
Ks
K.interrupt
Ks
Kf 1
K.configuration
Skp
(p-to-k; null / dual duplex)
Ks
Kf k

Pc/ central processor
Mp / primary memory
Ks / slow device control (e.g. for teletype)
Kf / fast device control (e.g. for disk)
Kc / control for clock, timer, interprocessor communications

Both switches have static configurations set by manual and program control

FIGURE ONE: C.mmp


P.host Unibus
High Impedence Probes
Interface
BM
BM
C
BM
EA
CR
BM
BM
C
BM
EA
CR
Primitive Event Detector
Accumulated Event Detector
Event Recorder (Data Buffer Control)
Interface
P.sup Unibus

BM / Bit mask
C / Comparitor
EA / Event accumulator
CR / Control register

FIGURE TWO: The C.mmp Hardware Monitor

ON HIGHER-ORDER SPECTRAL MOMENTS AND THE NORMALITY OF RANDOM PROCESSES: APPLICATION TO LINEAR AND NONLINEAR TRANSFORMATIONS

A. TRAGANITIS AND J. B. THOMAS
Department of Electrical Engineering
Princeton University, Princeton, N.J. 08540

ABSTRACT

We apply the theory of higher-order spectral moments to the problem of narrow-band filtering of random processes. We prove that the output converges to a Gaussian process. We pay particular attention to the rate of convergence and to methods to increase this rate. We show that band-pass filters are better behaved in this respect than low-pass filters of equal bandwidth. For a large class of processes we show that memoryless nonlinear transformations spread the output spectrum. In particular, this is true for Gaussian processes.

I. INTRODUCTION

The fact that narrow-band filtering of a random process tends to produce a Gaussian output is well known in the technical literature. Rigorous proofs have been presented by a number of authors [1,2]. They have usually assumed some kind of weak dependence for values of the input process which are temporally separated (for example, the strong mixing condition). After some manipulation of the output process, they obtain a series of independent r.v.'s to which the central limit theorem can be applied; hence the normality of the output. This approach, although mathematically attractive, shifts the attention to the time domain whereas it is more "natural" to work in the frequency domain as the name "narrow-band" indicates. In this paper we study this problem in the frequency domain and in doing so we use the notions of the "spectral moment" and "spectral semiinvariant" developed in [3].

Basic Definitions: The random process x(t) will be assumed to have zero mean and to belong to the classes S^∞ and Φ^∞, where, for $k=1,2,\ldots,x(t)\in S^\infty \Rightarrow E|x(t)|^k\le C_k<\infty$ and $Ex(t_1)\ldots x(t_k)=Ex(t_1+\tau)\times\ldots x(t_k+\tau),\forall\tau$ and $x(t)\in\Phi^\infty \Rightarrow$ there exist totally finite complex measures $M^k(\Lambda)$, the spectral moments, on the system of Borel sets B^k such that:

$$Ex(t_1)\ldots x(t_k)=m^k(t_1,\ldots t_k)=\int_{R^k} e^{i(\lambda,t)}M^k(d\lambda),\forall k$$

It is not hard to show that if $x(t)=\int_{R^1}e^{i\lambda t}X(d\lambda)$ is the spectral decomposition of x(t) then $M^k(d\lambda_1,\ldots,d\lambda_k)=EX(d\lambda_1)\ldots X(d\lambda_k)$. Since $m^k(t_1,\ldots,t_k)$ is a symmetric function of its arguments the measure $M^k(\Lambda)$ is symmetric. The semiinvariant $s^k(t_1,\ldots t_k)$ is defined as

$$s^k(t_1,\ldots,t_k) = \frac{\partial^k}{\partial a_1\ldots\partial a_k} \ln\{Ee^{i\Sigma a_j x(t_j)}\}_{||a||=0}$$

The spectral semiinvariant is a symmetric totally finite measure

$F^k(\Lambda)$ such that

$$s^k(t_1,\dots,t_k)=\int_{R^k} e^{i(\lambda,t)}F^k(d\lambda)$$

Because of stationarity, the spectral moments $M^k(d\lambda_1,\dots d\lambda_k)$ and the spectral semiinvariants $F^k(d\lambda_1,\dots d\lambda_k)$ are concentrated on the hyperplanes $\lambda_1+\lambda_2+\dots+\lambda_k=0$ of R^k. We can express the spectral moments as functions of the spectral semiinvariants and vice-versa with the following formulas:

$$F^k(d\lambda)=F^k(d\lambda_1,\dots d\lambda_k)=\sum(-1)^{m-1}(m-1)!M^{a_1}(d\lambda_{a_1})\dots M^{a_m}(d\lambda_{a_m}) \tag{1}$$

and

$$M^k(d\lambda)=\sum F^{a_1}(d\lambda_{a_1})\dots F^{a_m}(d\lambda_{a_m}) \tag{2}$$

where λ is the set $\{\lambda_1,\dots,\lambda_k\}$, and the summation is on all possible partitions of λ into non-ordered disjoint collections λ_{a_i}; a_i is the cardinality of λ_{a_i}. The characteristic functional (ch.f.) of a r.p. $x(t)$ is defined as:

$$\Phi_x(g)=E(e^{i\int x(t)g(dt)})=E(e^{i\int z(\lambda)X(d\lambda)})=\tilde{\Phi}_X(z) \tag{3}$$

where $g(t)$ is a real generalized totally finite measure on the Borel sets and $z(\lambda)$ its Fourier transform. The ch.f. admits a series expansion involving higher order spectral moments and its logarithm an expansion into spectral semiinvariants.

For a zero mean stationary Gaussian process it is easy to see that $F^k(\Lambda)=0, \forall k\neq 2$. For $k=2$, $F^2(d\lambda_1, d\lambda_2)=\delta(\lambda_1+\lambda_2)F(d\lambda_1)$ where $F(d\lambda)=EX(d\lambda)X(-d\lambda)$ is the spectral measure of the process. From Eq. (2), we get: $M^{2k+1}(\Lambda)=0, k=0,1,\dots$ and

$$M^{2k}(d\lambda_1,\dots d\lambda_{2k})=\sum \delta(\lambda_{i_1}+\lambda_{i_2})\dots\delta(\lambda_{i_{2k-1}}+\lambda_{i_{2k}})F(d\lambda_{i_1})\dots F(d\lambda_{i_{2k-1}}$$

where the summation extends over all possible pairs of λ_i's. The last expression implies that the even order spectral moments are concentrated on the intersection of the hyperplane $\lambda_1+\dots+\lambda_{2k}=0$ and the planes $\lambda_i+\lambda_j=0$. Therefore half of the arguments are on the positive real axis and the other half are symmetrically located on the negative axis.

II. THE NARROW-BAND FILTERING PROBLEM

In this section we study how a process becomes more nearly Gaussian as we restrict the bandwidth of its spectrum. Since the phenomenon depends more on the general bandwidth properties of the spectrum than on its exact shape, we will consider ideal narrow-band filters $H(\lambda)$; that is, $|H(\lambda)|=1$ for $|\lambda\pm\lambda_0|\leq\Delta$ and $|H(\lambda)|=0$ for $|\lambda\pm\lambda_0|>\Delta$. These filters although unrealizable (Wiener-Paley theorem) can be approximated to a satisfactory degree by realizable ones. In addition it will become clear that our arguments and conclusions remain valid for more realistic (complicated) filters at the expense of more complicated expressions.

Let $x(t)=\int_R e^{i\lambda t}X(d\lambda)$ be the input of a narrow-band filter

$H(\lambda)$. The normalized output is: $y(t)=\frac{1}{\sigma}\int_R e^{i\lambda t}H(\lambda)X(d\lambda)$ where $\sigma^2=\int_R|H(\lambda)|^2F(d\lambda)$. The spectral semiinvariants of y(t) are

$$F_y^k(L)=\frac{1}{\sigma^k}\int_L H(\lambda_1)\ldots H(\lambda_k)F_x^k(d\lambda_1,\ldots d\lambda_k)$$

The output will approach the Gaussian process iff $F_y^k(L)$ can be made to approach zero for any $k\neq 2$ and any $L\in R^k$. In our case this amounts to the requirement that $F_y^k(L)\to 0$ as the Lebesgue measure of L, $m(L)\to 0$. A sufficient condition is that $F_x^k(L)$ be absolutely continuous w.r. to the Lebesgue measure on the hyperplanes $\sum_1^k\lambda_i=0$. Processes of $S^\infty\cap\Phi^\infty$ with this property constitute the class Δ^∞ introduced by Kolmogorov [3]. The R-N derivative of $F_x^k(L)$ w.r. to m(L) is denoted by $f_x^k(\lambda_1,\ldots\lambda_k)$. Formally, we have $F_x^k(d\lambda_1,\ldots,d\lambda_k)=\delta(\sum_1^k\lambda_i)f_x^k(\lambda_1,\ldots,\lambda_k)d\lambda_1\ldots d\lambda_k$. We can now prove the following theorem:

Theorem 1. Let $x(t)\in\Delta^\infty$, $|f_x^k(\lambda_1,\ldots,\lambda_k)|\le A_k<\infty$. Let the power spectral density $f_x(\lambda)$ be bounded from below in the passband of the narrow-band filter $H(\lambda)$ so that $f_x(\lambda)\ge B>0$ for $|\lambda\pm\lambda_0|\le\Delta$. Then the output of the filter converges in probability to a Gaussian process as $\Delta\to 0$. Moreover the rate of convergence is greater for $\Delta<\lambda_0/3$.

Proof: Denote by Λ the support of the filter, $\Lambda=\{\lambda:|\lambda\pm\lambda_0|\le\Delta\}$; Λ^k is the k-fold direct product of Λ; $m(\Lambda)$ is the Lebesgue measure of Λ. The variance of the output is $\sigma^2=\int_R|H(\lambda)|^2F_x(d\lambda)\ge Bm(\Lambda)$

The ch.f. of y(t) is $\tilde{\Phi}_x(z.H)=E\exp[i\frac{1}{\sigma}\int_R z(\lambda)H(\lambda)X(d\lambda)]$ and by expanding its logarithm we have:

$$\log\tilde{\Phi}_x(z.H)=\sum_k\frac{1}{\sigma^k}\cdot\frac{i^k}{k!}\int_{R^k}z(\lambda_1)\ldots z(\lambda_k)H(\lambda_1)\ldots H(\lambda_k)F_x^k(d\lambda_1,\ldots d\lambda_k)$$

The first two terms of this expansion constitute the logarithm of the ch.f. of a Gaussian process; let us denote them by $\log\tilde{\Phi}_G(z.H)$. Therefore

$$|\log\tilde{\Phi}_x(z.H)-\log\tilde{\Phi}_G(z.H)|\le\sum_{k\ge3}\frac{1}{k!\sigma^k}\int_{\Lambda^{k-1}}|z(\lambda_1)|\ldots|z(\lambda_{k-1})|\times$$

$$|z(-d\lambda_1,\ldots,-d\lambda_{k-1})|\times|f_x^k(\lambda_1,\ldots,\lambda_{k-1},-\lambda_1-\ldots-\lambda_{k-1})|d\lambda_1\ldots d\lambda_{k-1}$$

For $z(\lambda)=\sum_1^n u_ie^{i\lambda t_i}$, $\tilde{\Phi}_x(z.H)=\Phi_{t_1,\ldots,t_n}(u_1,\ldots,u_n)$, which is the ordinary ch.f. of the output vector $(y(t_1),\ldots,y(t_n))$ and $\tilde{\Phi}_G(z.H)=\exp(-\frac{1}{2}U^TRU)$ where $R=\{\rho_{ij}\}=\{Ey(t_i)y(t_j)\}$ and $U^T=(u_1,\ldots,u_n)$. In this case $|\log\Phi_T(U)-(-\frac{1}{2}U^TRU)|\le\sum_{k\ge3}\frac{1}{k!\sigma^k}|U|^k\times A_k\,m(\Lambda)^{k-1}\le\sum_{k\ge3}\frac{1}{k!}|U|^k\frac{A_k}{B^{k/2}}m(\Lambda)^{(k-2)/2}$. It is clear that for $\Delta\to0$, $m(\Lambda)\to0$. Therefore $\Phi_T(U)\to\exp(-\frac{1}{2}U^TRU)$ which implies that the output vector $(y(t_1),\ldots,y(t_n))$ converges in probability to a Gaussian vector for any finite n; hence the convergence in probability to a Gaussian process.

Turning now to the problem of the rate of convergence, we see that, if we can make the third-order semiinvariant equal to zero, we will obtain a faster rate. This is exactly the case for $\Delta<\lambda_0/3$.

Lemma 1: If $\Delta<\lambda_0/3$, then $F_y^3(d\lambda_1,d\lambda_2,d\lambda_3)=0$

Proof: The proof amounts to showing the inconsistency of the restrictions imposed on the λ_i's, $\lambda_1+\lambda_2+\lambda_3=0$ and $|\lambda_i\pm\lambda_0|\leq\Delta$. Assume that $\lambda_1>0$ and $\lambda_2,\lambda_3\leq 0$. Then $\lambda_1\leq\lambda_0+\Delta$ and $\lambda_2+\lambda_3\leq-2\lambda_0+2\Delta$ Adding we get $0=\lambda_1+\lambda_2+\lambda_3\leq-\lambda_0+3\Delta$; i.e. $\lambda_0/3\leq\Delta$ is a contradiction.

From the above we deduce that

$$|\log\Phi_T(U)-(-\tfrac{1}{2}U^TRU)| \leq \begin{cases} \dfrac{A_3|U|^3}{6\,B^{3/2}}\,m(\Lambda)^{1/2}+o[m(\Lambda)^{1/2}]\ \text{if } \dfrac{\lambda_0}{3}\leq\Delta \\[2ex] \dfrac{A_4|U|^4}{24\,B^2}\,m(\Lambda)+O[m(\Lambda)]\ \text{if } \dfrac{\lambda_0}{3}>\Delta \end{cases}$$

It is clear therefore that a band-pass narrow band filter gives a faster rate of convergence than a low-pass filter ($\lambda_0=0$).

For the one-dimensional case we can apply the Berry-Esséen theorem to obtain a bound on the difference of the c.d.f. $F_y(x)$ of y(t) and the c.d.f. G(x) of the unit Gaussian r.v. (Kolmogorov distance). Ignoring the more rapidly decreasing terms we get

$$|F_y(x)-G(x)|\leq\frac{2}{\pi}\int_0^T\left|\frac{\Phi_y(u)-\exp(-\frac{1}{2}u^2)}{u}\right|du+\frac{1}{T(2\pi^3)^{1/2}}$$

$$\cong\frac{2}{\pi}\int_0^T\exp(-\tfrac{1}{2}u^2)\,\frac{1}{u}|\log\Phi_y(u)-(-\tfrac{1}{2}u^2)|\,du+\frac{3}{T}$$

For $\lambda_0/3\leq\Delta$ the dominating term in the bound is

$$\frac{2}{\pi}\,\frac{A_3}{6\,B^{3/2}}\,m(\Lambda)^{1/2}\int_0^T u^2\exp(-\tfrac{1}{2}u^2)du\leq 0.13\,\frac{A_3}{B^{3/2}}\,m(\Lambda)^{1/2}$$

For $\lambda_0/3>\Delta$ the dominating term is

$$\frac{2}{\pi}\,\frac{A_4}{24\,B^2}\,m(\Lambda)\int_0^T u^3\exp(-\tfrac{1}{2}u^2)du\leq 0.053\,\frac{A_4}{B^2}\,m(\Lambda)$$

Other means may be used to increase further the rate of convergence. After eliminating the third-order semiinvariant with the appropriate band-pass filter, the dominating semiinvariant is the fourth order. If it can be made zero by a ZNL in front of the filter, for example, the rate of convergence will be of the order of $m(\Lambda)^{3/2}$. If, in addition, $\Delta<\lambda_0/5$ then the fifth-order semiinvariant is also zero and the rate of convergence jumps by another $m(\Lambda)^{1/2}$ to $m(\Lambda)^2$.

We can gain further insight on the way a narrow band filter operates on the spectrum to produce a Gaussian output if we examine what happens to the higher order spectral moments. The assumption that x(t) is ergodic, although not strictly necessary for what follows, will simplify the presentation considerably. This is so because for ergodic processes $x(t)\in S^{\infty}\cap\Phi^{\infty}$, the spectral moments can be expressed by moments of smaller order through the relation [4]

$$\delta(\lambda_1+\ldots+\lambda_k)M^n(d\lambda_1,\ldots,d\lambda_n)=M^k(d\lambda_1,\ldots,d\lambda_k)M^{n-k}(d\lambda_{k+1},\ldots,d\lambda_n) \tag{5}$$

for any k=1,2,...,n-1. For $\lambda_1=-\lambda_{n+1}$, $\lambda_2=-\lambda_{n+2},\ldots,\lambda_n=-\lambda_{2n}$ we get

$$M^{2n}(d\lambda_1,\ldots,d\lambda_{2n})=M(d\lambda_1,-d\lambda_{n+1})\ldots M(d\lambda_n,-d\lambda_{2n})=F(d\lambda_1)\ldots F(d\lambda_n) \tag{6}$$

This implies that there is always present a Gaussian component in the higher order spectral moments of an ergodic process. It is exactly this component that passes through the filter whereas the rest of the spectral moments are attenuated.

Another nice property of ergodic processes which can be proved by using Eqs. (1) and (5) is that

$$\delta(\lambda_1+,\ldots,+\lambda_k)F^n(d\lambda_1,\ldots,d\lambda_n)=0 \text{ for } k\le n-1.$$

The relation

$$[1-\delta(\lambda_1+\ldots+\lambda_k)][M^n(d\lambda_1,\ldots,d\lambda_n)-F^n(d\lambda_1,\ldots,d\lambda_n)]=0 \tag{7}$$

for $k\le n-1$ is true in the general case.

Consider again an ergodic process x(t) and a filter $H(\lambda)$ satisfying the conditions of Theorem 1. The following two lemmas are true because of the restrictions imposed on the λ_i's, namely $\lambda_1+\ldots+\lambda_k=0$ and $|\lambda_i\pm\lambda_0|\le\Delta$. The proofs are by contradiction as in Lemma 1 and will not be given here.

Lemma 2: If $\lambda_0/\Delta>2k+1$ then $M_y^{2k+1}(L)=F_y^{2k+1}(L)=0, \forall L\in R^{2k+1}$.

Lemma 3: The even moments $M_y^{2k}(d\lambda_1,\ldots,\lambda_{2k})$ for $(2\lambda_0/\Delta)>2k$ have exactly k of their arguments on the positive axis.

It is now clear that as $\Delta\to 0$ an increasing number of the odd order spectral moments of the output will vanish. At the same time an increasing number of the even order moments will attain the desirable (Gaussian) property of Lemma 3. We may argue now that, in the limiting case, the λ_i's will be forced to form pairs of elements such that $\lambda_{i_1}+\lambda_{i_2}=0,\ldots,\lambda_{i_{2k-1}}+\lambda_{i_{2k}}=0$. Then, from Eq. (5) we obtain Eq. (4); that is, the even order moments become also the moments of a Gaussian process which implies convergence in probability of the output to a Gaussian process.

For non-zero bandwidths, however, there will be λ_i's which will not form pairs, with the result that additional terms will appear in the RHS of Eq. (4). These terms, through Eqs. (2), (5), (6), and (7), can be expressed by spectral semiinvariants of even (Lemma 2) order not less than 4. But, as we showed in Theorem 1, $F_x^k(\Lambda)\to 0$ as $m(\Lambda)\to 0$ faster than $F_x^2(\Lambda)$, and the greater the k the faster the convergence. Therefore the most significant of these additional terms are those which correspond to the situation where there are k-2 pairs of λ_i's and the remaining four λ_i's belong to an irreducible set, that is $\lambda_{i_1}+\lambda_{i_2}+\lambda_{i_3}+\lambda_{i_4}=0$ and $\lambda_{i_j}+\lambda_{i_\ell}\neq 0$ for any $j,\ell=1,2,3,4$. In this case, allowing the first 4 terms to belong to the irreducible set, we have

$$M_y^{2k}(\lambda_1,\dots,d\lambda_{2k})=F_x^4(d\lambda_1,\dots,d\lambda_4)F_x(d\lambda_5)F_x(d\lambda_7)\dots F_x(d\lambda_{2k-1})$$

Let us now calculate the difference due to these additional terms between the 2k-order moment of the output and the corresponding moment of a Gaussian process. We find

$$m^{2k}(t_1,\dots,t_{2k})=\frac{1}{\sigma^{2k}}\int_{R^{2k}} e^{i(\lambda,t)}H(\lambda_1)\dots H(\lambda_{2k})M_x^{2k}(d\lambda_1,\dots,d\lambda_{2k})$$

$$=\frac{1}{\sigma^{2k}}\sum_{*}\left[\int_{R^2}\exp[i(t_1\lambda_1+t_2\lambda_2)]\delta(\lambda_1+\lambda_2)H(\lambda_1)H(\lambda_2)F_x^2(d\lambda_1,d\lambda_2)\right]^{*k}$$

$$+\frac{1}{\sigma^{2k}}\sum_{**}\int_{R^4}\exp[i(\lambda_1t_1+\dots+\lambda_4t_4)]\delta(\lambda_1+\dots+\lambda_4)H(\lambda_1)\dots H(\lambda_4)F_x^4(d\lambda)\times$$

× [product of k-2 integrals like those appearing in the first term]

where * under the summation sign means summation over all possible ways of forming k pairs out of 2k λ_i's [(2k-1)(2k-3)...1 of them], *k means product of k terms of similar type, and ** means summation over all possible ways of forming k-2 pairs out of 2k λ_i's [(2k-1)(2k-3)...5 of them]. The first term in this equation turns out to be $\sum_{*}[Ey(t_i)y(t_j)]^{*k}$ which is the 2k-order moment of a normalized Gaussian process. The second term, which represents the deviation from normality, can be shown to be bounded by $(2k-1)(2k-3)\dots5\,\frac{A_4}{B^2}\,m(\Lambda)$ and therefore vanishes as the bandwidth decreases. The terms we ignored can be shown to be of $o[m(\Lambda)]$.

III. EFFECT OF ZNL'S ON SPECTRUM

Our goal in this section is to show that memory-less nonlinear transformations on Gaussian process have a "whitening" effect on the spectrum. Since this effect is true for other processes as well, we will study the problem in a more general setting, namely considering a second order stationary process x(t) whose bivariate p.d.f. admits a diagonal expansion into orthonormal polynomials. We write

$$p[x(t),x(t+\tau)]=p(x_1,x_2;\tau)=p(x_1)p(x_2)\sum a_n(\tau)\theta_n(x_1)\theta_n(x_2)$$

where $\{\theta_n(x)\}$ is a sequence of polynomials orthonormal w.r.t the weighting function p(x), the univariate p.d.f. For simplicity, we assume that Ex(t)=0.

Remark: The coefficient $a_n(\tau)$ is positive definite. To show that, let us consider a ZNL $\theta_n(\cdot)$. The autocorrelation $R_n(\tau)$ of $\theta_n[x(t)]$ is $R_n(\tau)=\iint\theta_n[x_1]\theta_n[x_2]p[x_1,x_2;\tau]\,dx_1dx_2=a_n(\tau)^n$. Therefore $a_n(\tau)$ is an autocorrelation function and as such its Fourier transform (if it exists) is nonnegative: $F[a_n(\tau)]=A_n(\lambda)\geqslant 0$. In particular, $A_1(\lambda)=[f_x(\lambda)]/\sigma^2$ where $f_x(\lambda)$ is the power spectral density of $x(t)$.

After passing $x(t)$ through a ZNL $g[\cdot]$ we obtain the process $y(t)=g[x(t)]$ whose autocorrelation is

$$R_y(\tau)=\iint g[x_1]g[x_2]p(x_1,x_2;\tau)dx_1dx_2=\sum a_n(\tau)\left[\int g[x]\theta_n(x)p(x)dx\right]^2=$$

$$=\sum a_n(\tau)c_n^2 \tag{8}$$

Transforming both sides, we get the power spectral density of $y(t)$

$$f_y(\lambda)=\sum c_n^2\,A_n(\lambda)=(c_1/\sigma)^2 f_x(\lambda)+\sum_{n>1} c_n^2 A_n(\lambda) \tag{9}$$

where we have ignored the dc term. The first term gives the contribution to the output spectrum due to the linear component of $g(\cdot)$. The remaining terms, which become more important for strong nonlinearities, indicate the contribution to the output spectrum due to the nonlinear components of $g(\cdot)$.

Notice that for any nonlinearity $g[\cdot]$ the output spectral components are completely determined by the output spectra $\{A_n(\lambda)\}$ of the ZNL's $\{\theta_n[\cdot]\}$ and only the relative contributions of these spectra are determined by the nonlinearity $g[\cdot]$. It is therefore sufficient to study polynomial type ZNL's in order to draw conclusions about the output spectrum of any ZNL. In the following we will focus our attention on one aspect of this problem; namely, the spreading of the spectrum due to the nonlinearity.

Let us first consider the important special case where $x(t)$ is a unit Gaussian process. Then [5], we have $\theta_n(x)=H_n(x)$ and $a_n(\tau)=[\rho^n(\tau)]/n!$ Therefore $A_n(\lambda)=[f_x^{*n}(\lambda)]/n!$ where $f_x^{*n}(\lambda)$ is the n-fold convolution of the input spectrum with itself. It is well known (central limit theorem) that the convolution operator smooths and spreads out the function. As a result $A_n(\lambda)$, and therefore the output spectrum of a ZNL, is distributed more widely along the axis than is the input spectrum. If we define the bandwidth as the normalized second moment of the spectrum; that is $BW=(\int\lambda^2 f(\lambda)d\lambda)/(\int f(\lambda)d\lambda)$, then from Eq. (9) we get

$$[BW]_y=[BW]_x\left\{\left[\sum c_k^2/(k-1)!\right]/\left[\sum c_k^2/k!\right]\right\}\geq[BW]_x \tag{10}$$

Some mild restrictions must be imposed on $g[\cdot]$ to allow the interchange of summation and intergration in Eq. (8). A sufficient condition for the validity of Eq. (8) and the subsequent conclusion in the Gaussian case is that $\int|g(x)|\exp(-x^2/4)dx<\infty$, which covers most of the practical ZNL's.

Other processes with the same properties [6] of the

$\{a_n(\tau)\}$ are those with bivariate Gamma distribution or, in particular, with exponential bivariate distribution. In both cases, $\theta_n(x)$ is the n-th degree Laguerre polynomial and $a_n(\tau)=\text{const.}\,\rho^n(\tau)$.

Another, more general, class of processes for which we can say something more about $A_n(\lambda)$, in addition to $A_n(\lambda)\geq 0$, is the class $S^\infty \cap \Phi^\infty$. For, if $x(t)\in S^\infty \cap \Phi^\infty$, then it is possible to express $A_n(\lambda)$ as a function of the spectral moments (or semiimvariants) of $x(t)$ of order not greater than 2n. Indeed let $\theta_n(x)=\sum_0^n b_{nk}x^k$. Then $a_n(\tau)=E\theta_n[x(t)]\theta_n[x(t+\tau)]=\sum_{kj}\sum b_{nk}b_{nj}Ex^k(t)x^j(t+\tau)$. But

$$\begin{aligned}
Ex^k(t)x^j(t+\tau) &= \int_{R^{k+j}} \exp\Big[it\sum_1^{k+j}\lambda_\ell + i\tau\sum_{k+1}^{k+j}\lambda_\ell\Big]M^{k+j}(d\lambda_1,\ldots,d\lambda_{k+j}) \\
&= \int_{R^{k+j}} \exp\Big[i\tau\sum_{k+1}^{k+j}\lambda_\ell\Big]\,\delta\Big[\sum_1^{k+j}\lambda_\ell\Big]M^{k+j}(d\lambda_1,\ldots,d\lambda_{k+j}) \\
&= \int_R e^{i\lambda\tau}\int_{R^{k+j-1}}\delta\Big(\sum_{k+1}^{k+j}\lambda_\ell-\lambda\Big)\delta\Big(\sum_1^{k+j}\lambda_\ell\Big)M^{k+j}(d\lambda_1,\ldots,d\lambda_{k+j}) \\
&= \int_R e^{i\lambda\tau}S^{k+j}(d\lambda)
\end{aligned} \tag{11}$$

Thus, not surprisingly, it is the higher order spectral moments of the input process which determine the output spectrum for this class. In the Gaussian case, we can express the spectral moments using the spectral measure as in Eq. (4); this enables us to express the output spectrum of a ZNL as a relatively simple function of the input spectrum as in Eq. (8). In the non-Gaussian case, such a simple relationship between higher-order spectral moments and spectral measure does not exist, and the matter is more complicated.

We can however draw some conclusions concerning the spreading of the spectrum. To illustrate, consider a band limited process $x(t)=\int_{-\Delta}^{\Delta}e^{i\lambda t}X(d\lambda)$ and one of the terms appearing in Eq. (11). To simplify the presentation, let k=j=2. Then

$$Ex^2(t)x^2(t+\tau)=\iiint_{-\Delta}^{\Delta}\exp[i(\lambda_3+\lambda_4)\tau]M^4(d\lambda_1,-d\lambda_1-d\lambda_3-d\lambda_4,d\lambda_3,d\lambda_4)$$

$$=\int_{-2\Delta}^{2\Delta}e^{i\lambda\tau}\iint_{-\Delta}^{\Delta}M^4(d\lambda_1,-d\lambda_1-d\lambda,d\lambda_3,d\lambda-d\lambda_3)=\int_{-2\Delta}^{2\Delta}e^{i\lambda\tau}S^4(d\lambda)$$

which implies that the Fourier transform of $Ex^2(t)x^2(t+\tau)$ may have non-zero values outside $(-\Delta,\Delta)$.

Concerning the smoothness of $S^4(L)$, we make the following observation. From Eq. (2) we can write:

$$M^4(d\lambda_1, -d\lambda_1-d\lambda, d\lambda_3, d\lambda-d\lambda_3)=$$

$$= F^2(d\lambda_1, -d\lambda-d\lambda_1)F^2(d\lambda_3, d\lambda-d\lambda_3)+F^2(d\lambda_1, d\lambda_3)F^2(-d\lambda-d\lambda_1, d\lambda-d\lambda_3)+$$

$$+ F^2(d\lambda_1, d\lambda-d\lambda_3) \quad F^2(d\lambda_3, -d\lambda-d\lambda_1)+R=\delta(\lambda)F(d\lambda_1)F(d\lambda_3)d\lambda +$$

$$+ \delta(\lambda_1+\lambda_3)F(d\lambda_1)F(d\lambda-d\lambda_3)+\delta(\lambda_1+\lambda-\lambda_3)F(d\lambda_1)F(d\lambda_3)+R$$

where R stands for the remaining terms in Eq. (2). Therefore $S^4(d\lambda)$ can be written as

$$S^4(d\lambda)=\delta(\lambda)d\lambda\iint F(d\lambda_1)F(d\lambda_3)+\iint\delta(\lambda_1+\lambda_3)F(d\lambda_1)F(d\lambda-d\lambda_3)+$$

$$+\iint\delta(\lambda_1+\lambda-\lambda_3)F(d\lambda_1)F(d\lambda_3)+\iint R = R_x^2(0)\delta(\lambda)d\lambda+2F^{*2}(d\lambda)+\iint R \qquad (12)$$

For a Gaussian process R=0, and the above equation gives the output spectrum of a square law ZNL [5]. The convolution in Eq. (12) is an indication of smoothness; the third term, which represents the deviation from normality, is the one that determines the smoothness in the final analysis. If it is small compared to the second term ("nearly Gaussian process") or if it behaves "nicely", then we can say that $S^4(L)$ is smoother than $F(L)$.

REMARKS

We have used the concepts of spectral moment and spectral semiinvariant to study some problems connected with the spectral properties of a process. Much remains to be done, however, concerning their further properties and their measurement. Finally we have pointed out a property of the linear and ZNL transformations which result in a Gaussian process: they both reduce the bandwidth.

ACKNOWLEDGEMENTS

This research is supported by the National Science Foundation under Grant GK-24187 and by the Air Force Office of Scientific Research under Grant AFOSR-71-2101.

REFERENCES

1. M. Rosenblatt, "A Central Limit Theorem and a Strong Mixing Condition", Proc. Nat. Acad. Sci. U.S.A., 42, 1956, pp. 43-47.

2. Yu. A. Rosanov, "Stationary Random Processes", Holden Day, 1967.

3. A. N. Shiryaev, "Some Problems in the Spectral Theory of Higher-order Moments", Theory of Probability and Its Applications, 5, 1960, pp. 265-284.

4. Ya. G. Siani, "On Higher-order Spectral Measures of Ergodic

Stationary Processes", Theory of Probability and Its Applications, 8, 1963, pp. 429-436.

5. J. B. Thomas, "An Introduction to Statistical Communication Theory", Wiley, New York, 1969.

6. K. V. Mardia, "Families of Bivariate Distributions", Griffin, London, 1970.

ESTIMATION OF A RANDOM SPACE-TIME FIELD THAT INFLUENCES AN OBSERVED SPACE-TIME POINT PROCESS*

PHILIP M. FISHMAN and DONALD L. SNYDER
Washington University, St. Louis, Missouri 63130

ABSTRACT

A space-time point process is a stochastic process having as realizations points with random coordinates in both space and time. We consider a general class of doubly-stochastic space-time point processes. These are space-time point processes influenced by an underlying and unobservable random space-time field, which we call the information process. The problem we consider is that of estimating the information process given observations of the point process. The conditional distribution of the information process is a function of the observations which is sufficient for forming estimates satisfying a wide variety of performance criteria. A general representation for this conditional distribution is given. This representation is the analog for observed space-time point processes of the Bucy representation associated with additive Gaussian observations.

I. PRELIMINARY DEFINITIONS

A stochastic point process on a Euclidean space X is a probability triplet $(\Omega,\mathcal{B},P)$ where the elements $\omega\varepsilon\Omega$ are denumerable point sets in X, $\mathcal{B}$ is the minimal σ-field generated by all sets of the form $\{\omega:N(\omega\cap A_i)=k_i,i=1,2,\ldots,j\}$ in which N(E) denotes the cardinality of the set E, A_i is a Borel set in X, k_i is a nonnegative integer, j is a positive integer, and P is any probability measure on $\mathcal{B}$. Each realization ω can be enumerated as $\omega=\{x_1,x_2,\ldots,x_j,\ldots\}$, where each x_i is a point in X. To give a natural ordering to the points, we now let X be the product space $[t_0,T)\times E^n$ where $-\infty<t_0<T\leq\infty$ and E^n is Euclidean n-space. Because in many applications $[t_0,T)$ is a time interval, we refer to point processes on such a state space as a *space-time point process*. Each realization ω of a space-time point process has the natural ordering $\{(t_1,\vec{r}_1),(t_2,\vec{r}_2),\ldots,(t_j,\vec{r}_j),\ldots\}$ where $t_0\leq t_i\leq t_{i+1}<T$ and $\vec{r}_i$ is a point in E^n for $i=1,2,\cdots$. We define the function

$$F_j[(\tau_1,\vec{\rho}_1),\ldots,(\tau_j,\vec{\rho}_j)]=\Pr[t_i\leq\tau_i,\vec{r}_i\leq\vec{\rho}_i;i=1,2,\ldots,j] \qquad (1)$$

where $\tau_i\varepsilon[t_0,T),\vec{\rho}_i\varepsilon E^n$ for $i=1,2,\ldots,j$ and $j\geq1$. Here the inequality $\vec{r}\leq\vec{\rho}$ means that each coordinate of $\vec{r}-\vec{\rho}$ is nonpositive. The function F_j is termed the j*th*-order distribution function of the space-time point process. Let $t\varepsilon[t_0,T)$ and denote by $\mathcal{B}_t$ the sub-σ-field of $\mathcal{B}$ generated by all sets of the form $\{\omega:N(\omega\cap A_i)=k_i;i=1,2,\ldots,j\}$ where now each A_i is a Borel set in $[t_0,t)\times E^n$. A space-time point process will be called *regular* if:

(a) the j*th*-order distribution function is an absolutely continuous function of its arguments for $j\geq1$;

(b) $$\Pr[N([t,\tau)\times E^n)>0|\mathcal{B}_t] < 1 \qquad \text{(a.s.) for all } \tau\varepsilon[t,T) \qquad (2)$$

(c) $$\lim_{\Delta t,\Delta r\downarrow0}\left\{\frac{\Pr[N([t,t+\Delta t)\times C(\vec{r},\Delta r)\geq2|\mathcal{B}_t]}{\Pr[N([t,t+\Delta t)\times C(\vec{r},\Delta r)\geq1|\mathcal{B}_t]}\right\}=0 \qquad \text{(a.s.)} \qquad (3)$$

where $C(\vec{r},\Delta r)$ denotes the n-dimensional cube $[r_1,r_1+\Delta r)\times\cdots\times[r_n,r_n+\Delta r)$.

Property (a) implies that a joint probability density exists for the occurrence locations of the point process. Property (b) implies that conditioned on the past, points do not occur with certainty in any future interval $[t,\tau)$. Property (c) is a generalization of Khinchin orderliness [1] which we call *conditional orderliness*.

---

*This work was supported in part by the National Institutes of Health under Research Grant RR 00396 from the Division of Research Facilities and Resources and by the National Science Foundation under Grant GK 32239.

II. PRELIMINARY RESULT

The following theorem is proven in [2].

THEOREM 1. Let $(\Omega,\mathcal{B},\mathcal{P})$ be a regular space-time point process. Then for almost every $\omega=\{(t_i,\vec{r}_i),i=1,2,\ldots\}$ there exists a $\mathcal{B}_t$-measurable function $\phi(t,\vec{r};\omega)$, which we call the *hazard function*, defined for $(t,\vec{r})\varepsilon[t_0,T)\times E^n$, such that:

(a) $\phi(t,\vec{r};\omega)$ is a Lebesque measurable function of t and $\vec{r}$

(b) Let D be a Borel set in E^n. Then for $j=0,1,2,\ldots$

$$\Pr[(t_{j+1},\vec{r}_{j+1})\varepsilon[t_j,t)\times D|\mathcal{B}_t](\omega)=\int_{t_j}^{t}\int_D\phi(\tau,\vec{r};\omega)\exp[-\int_{t_j}^{\tau}\int_{E^n}\phi(\xi,\vec{\rho};\omega)d\xi d^n\vec{\rho}]d\tau d^n\vec{r} \tag{4}$$

(c) $$\Pr[N([t,t+\Delta t)\times C(\vec{r},\Delta r))=1|\mathcal{B}_t](\omega)=\phi(t,\vec{r};\omega)\Delta t(\Delta r)^n + o[\Delta t(\Delta r)^n] \tag{5}$$

where $o[\Delta t(\Delta r)^n]/[\Delta t(\Delta r)^n]\to 0$ as $\Delta t,\Delta r\to 0$.

III. FURTHER DEFINITIONS

A doubly-stochastic space-time point process is defined as follows. Let $(\Omega_s,\mathcal{B}_s,\mathcal{P}_s)$ be a probability space. For each $\omega_s\varepsilon\Omega_s$, let there be a space time point process $(\Omega,\mathcal{B},\mathcal{P}(\cdot;\omega_s))$ on $[t_0,T)\times E^n$ such that for each $B\varepsilon\mathcal{B},\mathcal{P}(B;\cdot)$ is $\mathcal{B}_s$-measurable. Then a unique joint probability measure $\mathcal{P}'$ on the product space $(\Omega_s\times\Omega,\mathcal{B}_s\otimes\mathcal{B})$ can be constructed [3] such that

$$\mathcal{P}'(B_s\times B) = \int_{B_s}\mathcal{P}(B;\omega_s)\mathcal{P}_s(d\omega_s),\quad B_s\varepsilon\mathcal{B}_s,B\varepsilon\mathcal{B}. \tag{6}$$

A doubly-stochastic space-time point process is the space-time point process $(\Omega,\mathcal{B},\overline{\mathcal{P}})$ where $\overline{\mathcal{P}}(B)=\mathcal{P}'(\Omega_s\times B),B\varepsilon\mathcal{B}$. The realizations of $(\Omega,\mathcal{B},\overline{\mathcal{P}})$ may be considered as being obtained in the following manner. An element $\omega_s\varepsilon\Omega_s$ is selected randomly from $(\Omega_s,\mathcal{B}_s,\mathcal{P}_s)$. The point process realization ω is then obtained as a realization of $(\Omega,\mathcal{B},\mathcal{P}(\cdot;\omega_s))$. The probability $\overline{\mathcal{P}}(B)$ of an event $B\varepsilon\mathcal{B}$ is obtained by the averaging operation in (6) with $B_s=\Omega_s$.

It is shown in [2] that if $(\Omega,\mathcal{B},\mathcal{P}(\cdot;\omega_s))$ is regular for each ω_s, with corresponding hazard function $\phi(t,\vec{r};\omega,\omega_s)$, and if for each $B\varepsilon\mathcal{B}$ the conditional probability $\mathcal{P}(B;\omega_s|\mathcal{B}_t)(\omega)$ is $\mathcal{B}_s\otimes\mathcal{B}$-measurable, and if $\phi(t,\vec{r};\omega,\omega_s)$ is $\mathcal{P}'$-integrable, then $(\Omega,\mathcal{B},\overline{\mathcal{P}})$ is regular with hazard function $\hat{\phi}(t,\vec{r};\omega)=E[\phi(t,\vec{r};\omega,\omega_s)|(\Omega_s,\emptyset_s)\otimes\mathcal{B}_t](\omega)$, where $\emptyset_s$ denotes the null event in B_s.

IV. PROBLEM STATEMENT AND SOLUTION

The problem we consider is that of statistical inference about the underlying and unobservable realization ω_s which generates an observed realization of the space-time point process on $[t_0,t)\times E^n$. This problem is in principle solved by specifying the conditional probability measure $\mathcal{P}'[A\times\Omega|(\Omega_s,\emptyset_s)\otimes\mathcal{B}_t](\omega) \triangleq \mathcal{P}_s(A|\mathcal{B}_t)(\omega)$ for all $A\varepsilon\mathcal{B}_s$. It is shown in [2] that under the above conditions for the regularity of $(\Omega,\mathcal{B},\overline{\mathcal{P}})$ there holds

$$\mathcal{P}_s(A|\mathcal{B}_t)(\omega) = \frac{E_s\{\exp[H_t(\omega,\omega_s)]|\omega_s\varepsilon A\}\mathcal{P}_s(A)}{E_s\{\exp[H_t(\omega,\omega_s)]\}} \tag{7}$$

for all $A\varepsilon\mathcal{B}_s$ such that $\mathcal{P}_s(A)\neq 0$, where

$$H_t(\omega,\omega_s) = -\int_{t_0}^{t}\int_{E^n}\phi(\tau,\vec{\rho};\omega,\omega_s)d\tau d^n\vec{\rho} + \int_{t_0}^{t}\int_{E^n}\ell n\phi(\tau,\vec{\rho};\omega,\omega_s)N[\omega\cap(d\tau\times d^n\vec{\rho})]$$

and where the expectations in (7) are defined by $E_s[(\cdot)|\omega_s\varepsilon A]=\int_A(\cdot)\mathcal{P}_s(d\omega_s)/\mathcal{P}_s(A)$ and $E_s[(\cdot)]=\int_{\Omega_s}(\cdot)\mathcal{P}_s(d\omega_s)$. If $\mathcal{P}_s(A)=0$, then $\mathcal{P}_s(A|\mathcal{B}_t)=0$. We refer to (7) as the "Representation Theorem" for statistical inference for observed space-time point processes.

V. APPLICATIONS

1. Suppose $\phi(t,\vec{r};\omega,\omega_s)=\lambda(t,\vec{r},\underline{x}(t,\vec{r};\omega_s))$, where $\underline{x}$ is an essentially arbitrary space-time $\mathcal{B}_s$-measurable random field on $[t_0,T)\times E^n$. Then $(\Omega,\mathcal{B},\overline{\mathcal{P}})$ is a space-time generalization of the temporal doubly-stochastic Poisson process studied in [4]. Suppose an estimate of $\underline{x}(\tau,\vec{\rho};\omega_s)$ in terms of observations of the point process on $[t_0,t)\times E^n$ is desired, where $(\tau,\vec{\rho})\varepsilon[t_0,T)\times E^n$. If the prior density for $\underline{x}(\tau,\vec{\rho};\omega_s)$ exists, then a representation for the posterior density can be obtained from (7) by letting $A=\{\omega_s:\underline{X}\leq\underline{x}(\tau,\vec{\rho};\omega_s)<\underline{X}+\Delta\underline{X}\}$, dividing (7) by $\Pi_i\Delta X_i$, and taking the limit as $\Delta\underline{X}\to\underline{0}$. The result is

$$p(\underline{X};\tau,\vec{\rho}|\mathcal{B}_t)(\omega) = \frac{E_s\{\exp[H_t(\omega,\omega_s)]|\underline{x}(\tau,\vec{\rho};\omega_s)=X\}p(\underline{X};\tau,\vec{\rho})}{E_s\{\exp[H_t(\omega,\omega_s)]\}}. \tag{8}$$

This equation is the analog for observed space-time point processes of the Bucy Representation [5] when $\tau=t$ and the Generalized Bucy Representation [6] when $\tau\varepsilon[t_0,T)$. The use of the equation for solving filtering $(\tau=t)$, smoothing $(\tau<t)$, and prediction $(\tau>t)$ problems parallels the use in [5] and [6] of the alternative representations and is developed in [2].

2. Suppose $\phi(t,\vec{r};\omega,\omega_s)=\lambda(t,\vec{r},\omega;\underline{x}(t,\vec{r};\omega_s))$, where $\underline{x}$ is as before. Then $(\Omega,\mathcal{B},\overline{\mathcal{P}})$ is a space-time generalization of the temporal point processes studied in [7]. Equation (8) remains unchanged.

3. Suppose $\phi(t,\vec{r};\omega,\omega_s)=\lambda(t,\vec{r},\omega;\underline{x}(t,\vec{r};\omega,\omega_s))$, where $\underline{x}(t,\vec{r};\omega,\omega_s)$ is an essentially arbitrary space-time $\mathcal{B}_t\otimes\mathcal{B}_s$-measurable random field on $[t_0,T)\times E^n$. Then $(\Omega,\mathcal{B},\overline{\mathcal{P}})$ is a space-time generalization of the temporal point processes in [8] which have point-process dependent signals. Equation (8) remains unchanged.

4. Let $\phi(t,\vec{r};\omega,\omega_s)=\alpha(\omega_s)\lambda_1(t,\vec{r};\omega,\omega_s)+(1-\alpha(\omega_s))\lambda_0(t,\vec{r};\omega,\omega_s)$, where α is a $\mathcal{B}_s$-measurable random variable that has the value zero or one. Let $A_i=\{\omega_s:\alpha(\omega_s)=i\}$ and denote $\mathcal{P}_s(A_i)$ by $\mathcal{P}_i$, where $i=0,1$ and $\mathcal{P}_0+\mathcal{P}_1=1$. Then the likelihood ratio for making the decision $\alpha=1$ or $\alpha=0$ based on an observed realization of the space-time point process on $[t_0,t)\times E^n$ is given from (7) by

$$\frac{\mathcal{P}_s(A_1|\mathcal{B}_t)(\omega)}{\mathcal{P}_s(A_0|\mathcal{B}_t)(\omega)} = \frac{E_s\{\exp[H_t(\omega,\omega_s)]|\omega_s\varepsilon A_1\}\mathcal{P}_1}{E_s\{\exp[H_t(\omega,\omega_s)]|\omega_s\varepsilon A_0\}\mathcal{P}_0}. \tag{9}$$

The expectations required in (9) can be evaluated as

$$\begin{aligned}&E\{\exp[H_t(\omega,\omega_s)|\omega_s\varepsilon A_i\}\\&=\exp\{-\int_{t_0}^{t}\int_{E^n}\hat{\lambda}_i(\tau,\vec{\rho};\omega)d\tau d^n\vec{\rho}+\int_{t_0}^{t}\int_{E^n}\ell n\hat{\lambda}_i(\tau,\vec{\rho};\omega)N[\omega\cap(d\tau\times d^n\vec{\rho})]\}\end{aligned} \tag{10}$$

where

$$\begin{aligned}\hat{\lambda}_i(t,r;\omega) &\triangleq \frac{\int_{A_i}\lambda_i(t,\vec{r};\omega,\omega_s)\mathcal{P}_s(d\omega_s|\mathcal{B}_t)}{\mathcal{P}_s(A_i|\mathcal{B}_t)}\\ &= \frac{\int_{A_i}\lambda_i(t,\vec{r};\omega,\omega_s)\ \exp[H_t(\omega,\omega_s)]\mathcal{P}_s(d\omega_s)}{\int_{A_i}\ \exp[H_t(\omega,\omega_s)]\mathcal{P}_s(d\omega_s)}\end{aligned}$$

for $i=0,1$. Here, $\hat{\lambda}_i$ is a function of the observed point process on $[t_0,t)\times E^n$ which is the conditional mean of λ_i given the observations when $\alpha=i$. This result generalizes to space-time point processes the temporal Separation Theorem of Detection in [4].

5. Let $\underline{f}(t,\vec{r};\omega,\omega_s)$ be a $\mathcal{B}_s\otimes\mathcal{B}_t$-measurable random field on $[t_0,T)\times E^n$. It follows from (7) that the conditional-mean estimate $\hat{\underline{f}}$ defined by

$$\hat{\underline{f}} \triangleq \hat{\underline{f}}(t,\vec{r};\omega) \triangleq \int_{\Omega_s} \underline{f}(t,\vec{r};\omega,\omega_s)\mathcal{P}_s(d\omega_s|\mathcal{B}_t) \tag{11}$$

has the evaluation

$$\hat{\underline{f}} = \frac{\int_{\Omega_s} \underline{f}(t,\vec{r};\omega,\omega_s)\, \exp[H_t(\omega,\omega_s)]\mathcal{P}_s(d\omega_s)}{\int_{\Omega_s} \exp[H_t(\omega,\omega_s)]\mathcal{P}_s(d\omega_s)} .$$

We use this relation and the differential rule given in [9] to establish the following. Suppose $\underline{f}(t,\vec{r};\omega,\omega_s)=\underline{x}(t,\omega_s)$, where $\underline{x}$ satisfies the Ito differential-equation

$$d\underline{x}(t,\omega_s) = \underline{\alpha}(t,\omega_s)dt + \underline{\beta}(t,\omega_s)d\underline{W}(t,\omega_s), \quad \underline{x}(t_0,\omega_s) = \underline{x}_0(\omega_s) \tag{12}$$

in which $\{\underline{W}(t,\omega_s);t\geq t_0,\omega_s\varepsilon\Omega_s\}$ is a standard Wiener process and $\underline{\alpha}$ and $\underline{\beta}$ are $\mathcal{B}_{s\,t}$-measurable functions; here, $\mathcal{B}_{s\,t}$ is the sub-σ-field of $\mathcal{B}_s$ generated by the paths of $\underline{W}$ up to time t. Also, suppose the hazard function $\phi(t,\vec{r},\omega,\omega_s)$ is $\mathcal{B}_{s\,t}\otimes\mathcal{B}_t$-measurable. Then $\hat{\underline{x}}=\hat{\underline{x}}(t,\omega)$ satisfies

$$d\hat{\underline{x}} = [\hat{\underline{\alpha}}-\int_{E^n}(\widehat{\underline{x}\phi}-\hat{\underline{x}}\hat{\phi})d^n\vec{r}]dt + \int_{E^n}(\widehat{\underline{x}\phi}-\hat{\underline{x}}\hat{\phi})\hat{\phi}^{-1}N[\omega\cap(dt\times d\overset{n}{\vec{r}})] \tag{13}$$

with the initial condition $\hat{\underline{x}}(t_0,\omega)=E_s[\underline{x}_0(\omega_s)]$. This equation is used in [2] as the point of departure for developing a suboptimal estimator for $\underline{x}(t,\omega_s)$ when this process corresponds to the randomly changing position of a narrow beam of light that is incident on a spatially distributed photodetector.

Acknowledgement: The authors thank Prof. Thomas J. Kailath for bringing reference [9] to their attention.

REFERENCES

1. D. Daley and D. Vere-Jones, "A Summary of the Theory of Point Processes," in Stochastic Point Processes, P.A.W. Lewis (Ed.), p. 347, John Wiley and Sons, Inc., New York; 1972.
2. P. M. Fishman, "Statistical Inference for Space-Time Point Processes," Sc.D. Thesis, Washington University; to be submitted.
3. J. Neveu, The Calculus of Probability, p. 73, Holden-Day, Inc., San Francisco; 1965.
4. D. L. Snyder, "Filtering and Detection for Doubly-Stochastic Poisson Processes," IEEE Trans. Info. Th., IT-18, 1, pp. 91-102; Jan., 1972.
5. R. Bucy and P. Joseph, Filtering for Stochastic Processes with Applications to Guidance, p. 48, John Wiley and Sons, Inc., New York; 1968.
6. J. T. Lo, "On Optimal Nonlinear Estimation Part I: Continuous Observation," Info. Sciences, 6, pp. 19-32; 1973.
7. I. Rubin, "Regular Point Processes and Their Detection," IEEE Trans. on Info. Th., IT-18, 5, pp. 547-557; Sept., 1972.
8. A. Segall and T. Kailath, "Detection of Signals in Independent-Increment Type Noise," Proc. 1973 IEEE International Symp. on Information Theory, Ashkelon, Israel; 1973.
9. I Gihman and A. Dorogovcev, "On Stability of Solutions of Stochastic Differential Equations," Ukrain. Mat. Z., 6, pp. 229-250; 1965.

STATISTICAL PROPERTIES OF THE SCATTERING CHANNEL

LESTER F. EASTWOOD, JR.
Department of Electrical Engineering, and
Center for Development Technology
Washington University, St. Louis, Missouri 63130

ABSTRACT

Great current interest in communication theory centers on the statistical properties of the scattering (e. g. sonar, radar, and multipath) channel. The most general model for the random processes which arise in the scattering channel can account for the channel's detailed physical characteristics. However, relatively little has been discovered about the statistical properties of this model. By showing that the model is a linear process, this paper provides important information on these properties.

I. INTRODUCTION

This paper shows that the linear random process appropriately models some important random phenomena which arise in communication channels. To begin, we review the definition of the linear process and discuss some of its properties. With this work as a basis, we show that a general model [1] for reverberation in a scattering (reverberant) channel is a process of this type.

This fact contributes to the study of the reverberant channel problem. The linear processes are completely characterized statistically, and their properties have been intensively studied. The result of this paper shows that the powerful analytical tools developed for the linear process apply to the study of the scattering channel.

As one example of how this fact benefits the study of the problem, consider a common assumption which can now be made precise. After modeling reverberation noise, most detection theorists assume that it is Gaussian [2,3]. They base their assumption less on analysis than on an intuitive appeal to the Central Limit Theorem and on a desire for mathematical tractability. This lack of precision is dangerous, since reverberation noise measurements show the Gaussian assumption to be false in some situations [4]. We can discover when the Gaussian assumption is valid simply by using the existing analyses of conditions under which the general linear process is (or is nearly) Gaussian (e. g. [5]). These conditions, expressed in terms of easily measured properties of the channel, are easy to check in practice.

Modeling the reverberant channel has long been a popular topic for research. Classical works are the texts of Kennedy [2] and Van Trees [3]. Ol'shevskii [6] presents a summary of the efforts which have appeared on the sonar channel, while Middleton's series of papers [1] develop a general model of reverberation applicable to all scattering channels.

II. THE LINEAR PROCESS

In this section we present a general description of the linear process. This description is brief and is only for defining terminology; [7] contains a more careful exposition and lists further references.

We represent the linear process by stochastic integrals of the form

$$y(t) = \int_{R^n} f(t,\underline{s})\ dx(\underline{s}) \tag{1}$$

for all t in a set T and for R^n an arbitrary region in n-dimensional Euclidean space. The process $x(\underline{s})$ is a zero-mean, second order, independent increment process which is continuous in probability. In addition, for each t in T, $f(t,\underline{s})$ is real-valued and square integrable with respect to $dV(\underline{s}) = E|dx(\underline{s})|^2$ so that (1) is well-defined as a l. i. m. of Steiltjes sums [7]. Briefly stated, a linear process is an L_2 filtering, in the above sense, of an independent increment process.

The linear processes are completely characterized statistically, because we know their finite-dimensional characteristic functions (ch. f.'s). For the process in (1), these functions are

$$\Phi_{\underline{t}}(\underline{u}) = \exp\left\{\int_{R^n}\int_{-\infty}^{\infty} [\exp(ipw)-ipw-1]\quad (1/p^2)\ d_p d_{\underline{s}} G(p,\underline{s})\right\} \tag{2}$$

where $\underline{t} = t_1, \cdots, t_n$; $\underline{u} = u_1, \cdots, u_n$; and $w = u_1 f(t_1,\underline{s}) + \cdots + u_n f(t_n,\underline{s})$. The function $G(p,\underline{s})$ is monotone, nondecreasing in both p and $\underline{s}$, and is bounded in p. This function is uniquely determined by the ch. f. of an increment of the process $x(\underline{s})$ [7].

Knowledge of (2) allows us to evaluate any moment of the process y(t) simply by taking the appropriate derivative of the ch. f. Moveover, by investigating particular joint moments, we can discover much about the nature of the process. Theorems in [5,7], for example, specify conditions on the first few moments of y(t) which determine whether it is Gaussian, whether r. v.'s from the process are independent and whether a sequence of linear processes approaches a Gaussian process.

III. A GENERAL MODEL OF REVERBERATION

This section briefly describes Middleton's model [1] for the received waveform in a scattering channel and recognizes that this model is a linear process. Moreover, it discusses how this model may be generalized while retaining its complete statistical characterization. This latter discussion is of value in extending the applicability of the model.

To begin modeling the scattering channel, assume that the transmitter illuminates, and the receiver observes, a common region S of scatterers in space, as in Fig. 1.

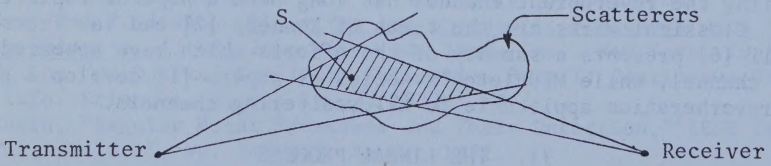


Fig. 1 A Transmitter and Receiver with a Common Illuminated Region S of Scatterers in Space

Let the illuminated scatterers have random spatial coordinates $\underline{s}_i$, and assume that the $\underline{s}_i$ are independent and that the probability of two or more points being in the same location is zero. These assumptions imply that these points have an inhomogeneous Poisson distribution in space. Associate with each scatterer a set of random parameters $\underline{\theta}_i \in \Theta$, letting each $\underline{\theta}_i$ depend only on its location $\underline{s}_i$ so that the $\underline{\theta}_i$ are mutually independent. Assume that the reflected waveform from the i-th scatterer equals $U(t;\underline{s}_i,\underline{\theta}_i)$ at the receiver at time t and that all reflected waveforms share this functional form. Assuming linearity of the channel, the superposition of these waveforms is the signal at the receiver, y(t). Middleton shows that y(t) can be written as (see [1], pp. 376-7):

$$y(t) = \int_S U(t;\underline{s},\underline{\theta})\, dx(\underline{s},\underline{\theta}) \tag{3}$$

where $x(\underline{s},\underline{\theta})$ is a marked Poisson process on S with random parameters $\underline{\theta} \in \Theta$. Eq. (3) has the form of the linear process in (1), so y(t) in (3) is a linear process. As verification of this fact, the ch. f. derived in [1] for (3) has the form of (2).

This model's assumptions may be generalized while retaining its complete statistical characterization. For example, the ch. f. of the sum of the process in (3) and an independent Gaussian process is easy to find; in fact, if we allow $x(\underline{s},\underline{\theta})$ to have a Brownian motion component, the form of the ch. f. (2) for y(t) in (3) is unchanged. Moveover, Middleton assumes that the $\underline{\theta}_i$ are identically distributed for all i, but this assumption is unnecessary for (3) to be a linear process with ch. f. (2)

IV. APPLICATIONS OF THE RESULT

This paper recognizes that the existing results on linear processes apply in scattering channel analysis. This information, in combination with existing work on detection and estimation for linear processes [8,9], can be used to design optimum communication systems for radar, sonar, and multipath channels. In addition, as stated earlier, the information is useful in verifying independence and Gaussianity assumptions.

ACKNOWLEDGEMENT

This work was financed by the National Science Foundation, under Grant GK-24187, and by the U. S. Army Research Office, under Contract DAHCO4-69-C-0012.

REFERENCES

1. D. Middleton, "A Statistical Theory of Reverberation and Similar First-Order Scattered Fields," IEEE Trans. Info. Theory, Vol. IT-13, No. 3, July 1967, pp. 372-92.

2. R. S. Kennedy, Fading Dispersive Communication Channels, New York: Wiley, 1969.

3. H. L. Van Trees, Detection, Estimation, and Modulation Theory, Part III, New York: Wiley, 1971.

4. G. V. Trunk and S. F. George, "Detection of Targets in Non-Gaussian Sea Clutter," IEEE Trans, on Aero, and Elec. Syst., Vol. AES-6, No. 5, September 1970, pp. 620-28.

5. R. Lugannani and J. B. Thomas, "The Central Limit Theorem for a Class of Stochastic Processes," J. Math. Anal. and Appl., Vol. 24, No. 1, Oct. 1968, pp. 25-38.

6. V. V. Ol'shevskii, Characteristics of Sea Reverberation, Consultants Bureau; New York, 1967.

7. R. Lugannani and J. B. Thomas, "On a Class of Stochastic Processes which are Closed under Linear Transformations," Info. and Control, Vol. 10, No. 1, Jan. 1967, pp. 1-21.

8. L. F. Eastwood and R. Lugannani, "Approximately Optimum Detectors for Linear Processes," Proc. Tenth Annual Allerton Conference on Circuit and System Theory, Univ. of Ill: Urbana, 1972.

9. L. F. Eastwood, Detection Theory Applied to Linear Processes, Ph.D. Thesis, Princeton University, 1973.

ON A RELATIONSHIP BETWEEN LIKELIHOOD RATIOS AND CONDITIONAL MEAN ESTIMATES FOR THE EXPONENTIAL FAMILY OF DISTRIBUTIONS

S. C. SCHWARTZ*
Department of Electrical Engineering
Princeton University, Princeton, N.J. 08540

ABSTRACT

A relationship is derived between likelihood ratios and conditional mean estimates for the exponential family of distributions. It is shown that the logarithm of the likelihood ratio is given by the correlation between this estimate and the differential of the sufficient statistic which characterizes the exponential family. This result extends the "estimator-correlator" receiver discussed in the literature for Gaussian noise. A canonical form for the a posteriori density is derived and the relationship between likelihood ratio structure and prior distribution is discussed through examples.

I. INTRODUCTION

In this study, we investigate the structure and the relationship between the average likelihood ratio and the conditional mean estimate (CME) for the exponential family of distributions.[1] Given a random vector $\underline{x}$, whose distribution depends on a random variable θ, the simplest subclass of the exponential family can be written in the form:

$$p(\underline{x}/\theta) = K(\theta)g(\underline{x})\exp(s(\theta)k(\underline{x})). \tag{1.1}$$

From the factorization theorem [1], it follows that $k(\underline{x})$ is a sufficient statistic of dimension one. It is not surprising that in any detection or estimation problem associated with (1), likelihood ratios and conditional mean estimates can be expressed in terms of the sufficient statistic. In fact, we will show that the structure of the likelihood ratio is uniquely determined by the CME which, in turn, is determined once the prior probability on the parameter θ is specified.

For a binary detection problem associated with (1), assume the conditional likelihood ratio takes the form:

$$\ell(\underline{x}/\theta) = \frac{p_1(\underline{x}/\theta)}{p_o(\underline{x}/\theta_o)} = K(\theta)e^{s(\theta)k(\underline{x})} \tag{1.2}$$

Then, the average likelihood ratio is shown to be given by

$$\ell(\underline{x}) = \int \ell(\underline{x}/\theta)\pi(\theta)d\theta = c_1 \exp\{\int \hat{s}(x)dk(x)\} \tag{1.3}$$

where $\hat{s}(\underline{x})$ is the conditional mean estimate

---

*Research reported here was supported in part by the John Simon Guggenheim Memorial Foundation and by the National Science Foundation under Grant GK-35359.

[1] This family includes, among others, the Gaussian, gamma, Poisson and beta.

$$\hat{s}(\underline{x}) = \int_\theta s(\theta) p_1(\theta/\underline{x}) d\theta \tag{1.4}$$

and c_1 is a constant which may depend on the number of observations (the dimension of $\underline{x}$) and underlying statistics, but which is independent of the actual data $\underline{x}$.

The likelihood ratio is thus seen to be a function of the correlation between the CME and the sufficient statistic. This result is a generalization of the "estimator-correlator" receiver often discussed in the literature under a Gaussian noise assumption, i.e. $p_1(x/\theta)$ is normal. It is also related to a result of Esposito [2] who showed that under mild regularity conditions, the average likelihood ratio could be written in terms of the conditional likelihood ratio and a pseudo-estimate $\tilde{s}$:

$$\ell(\underline{x}) = \ell(\underline{x}/\tilde{s}(\underline{x})) \tag{1.5}$$

$\tilde{s}$ is called a pseudo-estimate because it may not be a good estimate for estimation purposes; it is the estimate required for optimum detection. Esposito then shows that for $p(\underline{x}/\theta)$ Gaussian, the pseudo-estimate is equal to the CME, $\tilde{s}(\underline{x}) = \hat{s}(\underline{x})$. Equation (3) above indicates that this equality of estimates is true for all members of the exponential family.

The canonical structure (3) and the related a posteriori density $p(\theta/\underline{x})$ allow us to answer questions like: does there exist an a priori density such that the likelihood ratio is of a specified form[2]; is there an asymptotic form for the likelihood ratio which is independent of the particular prior distribution $\pi(\theta)$. Examples will be given to illustrate these points and will include a standard and non-standard Gaussian problem, and gamma statistics.

Generalizations discussed will include: the exponential family with sufficient statistics of dimension greater than one; the case when θ is a random vector; a recursive formulation for the likelihood ratio.

STRUCTURE OF THE LIKELIHOOD RATIO

In the sequel, we drop the vector notation and write $\underline{x} = x$. As discussed in the introduction, we take the conditional likelihood ratio in the form

$$\ell(x/\theta) = K(\theta)g(x)e^{s(\theta)k(x)} \tag{2.1}$$

The total differential of $\ell(x/\theta)$ is:[3]

---

[2] This is related to a study of Balakrishnan who investigated processes for which minimum mean-square estimates are of a specified form [3].

[3] We use differentials in contrast to gradients so as to handle discrete as well as continuous random variables.

$$d\ell(x/\theta) = \sum_{i=1}^{N} \frac{\partial \ell(x/\theta)}{\partial x_i} dx_i$$

$$= \ell(x/\theta)s(\theta) \sum_{i=1}^{N} \frac{\partial k}{\partial x_i} dx_i + K(\theta)e^{s(\theta)k(x)} \sum_i \frac{\partial g}{\partial x_i} dx_i$$

$$= \ell(x/\theta)\left\{s(\theta)dk(x) + \frac{1}{g(k)} \; dg(x)\right\} \tag{2.2}$$

The average (or generalized) likelihood ratio (ALR) is defined by

$$\ell(x) = \frac{p_1(x)}{p_o(x)} = \int_\theta \ell(x/\theta)d\pi(\theta) \tag{2.3}$$

Assuming $d\ell(x/\theta)$ is integrable with respect to $d\pi(\theta)$, we have for the differential of $\ell(x)$:

$$d\ell(x) = \int_\theta d\ell(x/\theta)\, d\pi(\theta)$$

$$= \int_\theta \{s(\theta)dk(x) + \frac{dg(x)}{g(x)}\} \ell(x/\theta)d\pi(\theta) \tag{2.4}$$

Use the following form of Bayes rule

$$\ell(x/\theta)d\pi(\theta) = \ell(x)dP(\theta/x) \tag{2.5}$$

to obtain

$$d\ell(x) = \ell(x)\{\hat{s}(x)dk(x) + \frac{dg(x)}{g(x)}\} \tag{2.6}$$

where $\hat{s}(x)$ is the C.M.E.

$$\hat{s}(x) = \int_\theta s(\theta)dP(\theta/x) \tag{2.7}$$

Dividing by $\ell(x)$ and integrating gives the desired result.

Theorem 1. Assume $\ell(x/\theta)$ is given by (2.1), $\ell(x) \neq 0$, and that $d\ell(x/\theta)$ is integrable with respect to $\pi(\theta)$. Then, the average likelihood ratio is

$$\ell(x) = \frac{1}{c_1} g(x) \exp\{\int \hat{s}(x)dk(x)\} \tag{2.8}$$

with the normalizing constant

$$c_1 = E_{H_o}\{g(x) \exp\{\int \hat{s}(x)dk(x)\}\}$$

$$= \int g(x) \exp\{\hat{s}(x)dk(x)\}p_o(x)dx \tag{2.9}$$

Observe that the constant does not depend on the observation vector x, but only on the underlying statistics and the number of observations. Consequently, when $g(x)=1$, as is the case in many detection problems, the logarithm of the ALR is a correlation between the sufficient statistic and the CME.

From (2.5), we also have an explicit expression for the a posteriori distribution.

$$dP(\theta/x) = c_1 \; K(\theta) \; d\pi(\theta) \; \exp\{\int (s(\theta) - \hat{s}(x))dk(x)\} \tag{2.10}$$

Hence, the a posteriori distribution is a function of the correlation between the error in the estimate and the statistic $k(x)$. Since (2.10) integrates to one, we also obtain an integral equation which the prior distribution must satisfy:

$$\exp\{\int \hat{s}(x)dk(x)\} = c_1 \int_\theta K(\theta) e^{s(\theta)k(x)} \tag{2.11}$$

This relationship will be useful later when we investigate in more detail the structure of the ALR and its relationship to the CME and possible prior distributions.

Alternate expressions for $\ell(x)$ and $dP(\theta/x)$ can easily be obtained. Observe that x enters the ratio $\ell(x/\theta)/\ell(x)$ only through the sufficient statistic $k(x)$. It then follows from (2.5) that $dP(\theta/x)$ and the CME of $s(\theta)$ are a function of x only through $k(x)$. Thus, we may write

$$s(x) = \int s(\theta)dP(\theta/x)dx = s(k(x)) \tag{2.12}$$

Defining $F(k) = \int f(k)dk$, the logorithm of the ALR becomes

$$L(x) = \ln(\ell(x)) = F(k) - \ln c_1 \tag{2.13}$$

Suppose we specify a prior distribution such that the CME is of the form $\hat{s}(x) = c_2/k(x)$. It immediately follows that the ALR is $\ell(x) = \frac{(k(x))^{c_2}}{c_1}$. Consequently, we can formulate and (sometimes) answer questions like: does there exist an a priori distribution such that the ALR (or equivalently the CME) is of a specified form.

The examples in the next section illustrate some of the applications of the above results.

EXAMPLES

Example 1 - The Gaussian Family

In this example we consider estimates which are linear functions of the sufficient statistic,

$$\hat{s}(x) = c_2 \; k(x) \tag{3.1}$$

From (2.8), the ALR becomes

$$\ell(x) = \frac{1}{c_1} g(x) \; \exp\{c_2 \frac{k^2(x)}{2}\} \tag{3.2}$$

Suppose θ is defined on the real line, $-\infty<\theta<\infty$, and that $s(\theta)=\theta$. Let $k(x)=iu$, where $i=\sqrt{-1}$. Then, utilizing (2.11), we obtain

$$\exp\left\{\frac{-c_2 u^2}{2}\right\} = c_1 \int_{-\infty}^{+\infty} K(\theta) e^{i\theta u} d\pi(\theta) \tag{3.3}$$

We recognize the left side of this equation as the Fourier transform of a Gaussian density function with zero mean and variance c_2. It follows from the uniqueness of the transform that

$$c_1 \; K(\theta)\pi(\theta) = \frac{1}{\sqrt{2\pi \; c_2}} \; e^{\frac{\theta^2}{-2c_2}} \tag{3.4}$$

Let us now assume that the conditional density $p_1(\underline{x}/\theta)$ is Gaussian; i.e., the observation vector is $\underline{x}=\theta+\underline{\omega}$, where $\underline{\omega}$ is a zero-mean Gaussian noise:

$$p_1(\underline{x}/\theta)=(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i-\theta)^2\right) \tag{3.5}$$

Under hypothesis H_o, we can assume $p_o(x)$ is of the same form with $\theta=0$. It then follows that

$$\ell(x/\theta)=\exp\left\{\frac{-N\theta^2}{2\sigma^2} + \frac{\theta}{\sigma^2} \sum_{i=1}^{N} x_i\right\}$$

Identifying terms, we have:

$$K(\theta) = \exp\left\{\frac{-N\theta^2}{2\sigma^2}\right\}, \quad g(x) = 1$$

$$k(x) = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i, \quad s(\theta)=\theta \tag{3.6}$$

From (3.4) and (3.6), we see that $\pi(\theta)$ is uniquely determined and must be of the form

$$\pi(\theta) = \frac{1}{\sqrt{2\pi c_2 c_1^2}} \exp\left\{\frac{-\theta^2}{2c_2} + \frac{N\theta^2}{2\sigma^2}\right\} \tag{3.7}$$

Carrying the details further, we have

$$c_1^2 = \frac{\sigma^2+N\sigma_\theta^2}{\sigma^2} = \frac{\sigma_\theta^2}{c_2}, \quad c_2 = \frac{\sigma_\theta^2\sigma^2}{\sigma^2 + N\sigma_\theta^2} \tag{3.8}$$

and $\pi(\theta)$ is the Gaussian density with standard deviation σ_θ. It is a simple matter to verify these results by calculating the a posteriori density from (3.5) and (3.7). It is

$$p(\theta/\underline{x}) = (2\pi c_2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2c_2} (\theta-c_2 \; k(\underline{x}))^2\right\}$$

The conditional mean estimate is $\hat{\theta} = c_2\, k(\underline{x})$, which coincides with our initial assumption.

It is interesting to note that the assumption of a linear estimate, Eq. (3.1), leads to a Gaussian posterior density. Then, having specified $p(\underline{x}/\theta)$ as Gaussian uniquely determines $\pi(\theta)$ to be of the same type. One can define a family of densities $p(x/\theta)$ for which (3.4) is satisfied, the posterior density is Gaussian and, yet, the prior is non-Gaussian. (Clearly, this is not the usual additive noise problem.) Consider the conditional density

$$p_1(x/\theta) = \frac{1}{\sigma^2+\theta^2}\ \frac{x^2}{\sqrt{2\pi}\ \sigma}\ e^{-(x-\theta)^2/2\sigma^2} \tag{3.11}$$

We take $p_o(x)$ equal to $p_1(x/\theta)$ with $\theta=0$. Identifying terms:

$$\ell(x/\theta) = \frac{\sigma^2}{\sigma^2+\theta^2}\exp\left\{-\frac{\theta^2}{2\sigma^2} + \frac{x\theta}{\sigma^2}\right\}$$

$$k(x) = \frac{x}{\sigma^2}\ ,\quad s(\theta)=\theta$$

$$K(\theta) = \frac{\sigma^2}{\sigma^2+\theta^2}\ \exp\left\{-\frac{\theta^2}{2\sigma^2}\right\}$$

The a priori density, which is similar to a Gaussian mixture, is uniquely determined to be

$$\pi(\theta) = \frac{(\sigma^2+\theta^2)}{(\sigma^2+\sigma_\theta^2)}\ \frac{1}{\sqrt{2\pi}\ \sigma_\theta}\ \exp\left\{-\frac{\theta^2}{2\sigma_\theta^2}\right\} \tag{3.12}$$

This prior satisfies (3.4) with c_2 given as before. Hence the CME is identical to (3.1)

Example 2 - Renewal Process with Gamma Statistics

Let $\{x_i\}_1^j = \{x_1, x_2, \ldots x_j\}$ denote the inter-arrival times and suppose that under the two hypotheses we have

$$H_1:\ p_1(x_i/\theta_1) = \frac{\theta_1^{\beta}}{\Gamma(\beta)}\ x_i^{\beta-1}\ e^{-\theta_1 x_i};\ \theta_1 > 0 \tag{3.13}$$

$$H_0:\ p_0(x_i/\theta_0) = p_1(x_i/\theta_0)$$

The conditional likelihood ratio is

$$\ell(x/\theta) = \left(\frac{\theta_1}{\theta_0}\right)^{j\beta}\ \exp\{-(\theta_1-\theta_0)\sum_{i=1}^{j} x_i\} \tag{3.14}$$

and the various terms are:

$$g(x)=1 \ , \ k(x) = -\sum_{i=1}^{j} x_i$$

$$K(\theta) = \left(\frac{\theta_1}{\theta_o}\right)^{j\beta} \quad , \quad s(\theta) = (\theta_1 - \theta_o) \tag{3.15}$$

From our previous results, the logarithm of the average likelihood ratio is simply

$$\ln \ell(x) = \int E\left\{\theta_1 - \theta_0 \mid \{x_i\}_1^j\right\} dk(x) \tag{3.16}$$

Now, it turns out that the maximum likelihood estimate for θ_1, $\max\limits_{\theta_1} p_1(x/\theta_1)$, is

$$\hat{\theta}_{1(ML)} = \frac{j\beta}{\sum\limits_{i=1}^{j} x_i} = \frac{j\beta}{-k} = \frac{\beta(\#\text{ of counts})}{\text{time}} \tag{3.17}$$

Suppose we ask if there exists a prior density such that the CME is also of the same form, i.e., the estimate is inversely proportional to the sufficient statistic:

$$E\{(\theta_1 - \theta_0) \mid \{x_i\}_i^j\} = \hat{s}(x) = \frac{c_2}{k(x)} + c_3 \tag{3.18}$$

Using the results of section 2, we obtain

$$\ell(x) = \exp\{\int \hat{s}(x)\,dk(x)\} = c_1 \int_0^\infty K(\theta)\pi(\theta_1) e^{s(\theta)k(x)} d\theta_1$$

$$k^{c_2} e^{c_3 k} = c_1 \int_0^\infty K(\theta)\pi(\theta_1) e^{(\theta_1 - \theta_0)k} d\theta_1 \tag{3.19}$$

If we set $c_3 = -\theta_0$, we have

$$k^{c_2} = c_1 \int_0^\infty K(\theta)\pi(\theta_1) e^{\theta_1 k} d\theta_1 \tag{3.20}$$

Now identify k as a Laplace variable, and let $c_2 = -N$, where N is a positive integer. We uniquely determine $K(\theta)\pi(\theta_1)$ to be the inverse Laplace transform of $1/k^N$:

$$K(\theta)\pi(\theta_1) = \frac{\theta_1^{N-1}}{c_1 (N-1)!} \tag{3.21}$$

and the prior is

$$\pi(\theta_1) = \frac{\theta_1^{N-1}}{c_1 (N-1)!} \left(\frac{\theta_0}{\theta_1}\right)^{j\beta} , \ \theta_1 > 0 \tag{3.22}$$

For the CME and the likelihood estimate to coincide, we require $N=j\beta$. Clearly, this is not possible (both mathematically and physically) since $\pi(\theta_1)$ is not integrable. It is possible, however, to have a CME which is inversely proportional to the sufficient statistic. Assuming hypothesis H_1 represents a signal plus noise situation (e.g., $\theta_1=\theta_0+\theta_1$) with $j\beta>N$, the prior density $\pi(\theta)$ is well defined with finite moments up to order r, where r is the largest integer which satisfies $r<j\beta-N$.

The above discussion is for any gamma density with general parameter β. Let us now assume that $\beta=2$ and that we have one observation. The prior is

$$\pi(\theta_1)=\frac{(\theta_0+\theta_1)^{N-1}}{c_1(N-1)!}\left(\frac{\theta_0}{\theta_0+\theta_1}\right)^2 \tag{3.23}$$

For an integrable density, we require N=1. It then follows that

$$\hat{s}(x) = \frac{-1}{k(x)} - \theta_0 = \frac{1}{x} - \theta_0 \tag{3.24}$$

$$\pi(\theta_1) = \frac{\theta_0^2}{c_1} \frac{1}{(\theta_0+\theta_1)^2}$$

$$= \frac{\theta_0}{(\theta_0+\theta_1)^2} \tag{3.25}$$

Using this prior and (3.13) in the form

$$p_1(x/\theta_1) = (\theta_1+\theta_0)^2 \, x \, e^{-x(\theta_0+\theta_1)},$$

it is easy to obtain the a posteriori density directly by means of Bayes rule:

$$p(\theta_1|x) = x\,e^{-\theta_1 x} \tag{3.26}$$

which is the exponential density (gamma with $\beta=1$). Alternatively, one could derive this a posteriori density by making the appropriate substitutions in the canonical structure, Eq. (2.10).

Clearly, one could assume other forms for the CME and determine whether a prior exists. Perhaps the more important question is the structure of the likelihood ratio. In the above example, with $\hat{s}(x)$ of the form (3.18), the active part of the logarithm of the ALR is

$$k(x)+\frac{c_2}{c_3}\ln k(x)$$

which, for a large number of counts, $k(x)= -\sum_{i=1}^{j} x_i$, is approximated by just the counts.

4. GENERALIZATIONS

The previous results can be extended in a number of directions. These include: the exponential family with sufficient statistics of dimension greater than one; the situation where

θ is a random vector; a recursive formulation for the likelihood ratio.

For members of the exponential family given by

$$p_1(\underline{x}/\theta)=K(\theta)g(\underline{x})\exp\left\{\sum_{i=1}^{M} s_i(\theta)k_i(\underline{x})\right\} \tag{4.1}$$

$k(x)=\{k_1(x),k_2(x),\ldots,k_M(x)\}$ is a sufficient statistic of dimension M([1]). When the conditional likelihood ratio is of the same form, the following results are obtained in a manner completely analogous to that of section 2.

The ALR is

$$\ell(\underline{x})=\frac{1}{c_1}g(x)\exp\left\{\sum_{i=1}^{M}\int\hat{s}_i\,dk_i(x)\right\}, \tag{4.2}$$

with the constant c_1 calculated as before (see (2.9)).

The a posteriori distribution becomes

$$dP(\theta/x) = c_1K(\theta)d\pi(\theta)\exp\left\{\sum_{i=1}^{M}\int(s_i(\theta)-\hat{s}_i(x))dk_i(x)\right\} \tag{4.3}$$

where

$$\hat{s}_i(x) = \int_\theta s_i(\theta)dP(\theta/\underline{x}) \tag{4.4}$$

The examples in section 3 then have direct extensions for the vector parameter case.

Another generalization is the case where $\underline{s}(\underline{\theta})$ is of the same dimension as the observation vector $\underline{x}$. Assume a likelihood ratio of the form

$$\begin{aligned}\ell(\underline{x}/\underline{\theta}) &= K(\theta)\exp\{\underline{k}(x)R\underline{s}(\underline{\theta})\} \\ &= K(\theta)\exp\{\sum_{i,j}^{N} k_i(\underline{x})r_{ij}s_j(\underline{\theta})\}\end{aligned} \tag{4.5}$$

where R is a positive definite matrix. Define the vector of conditional mean estimates

$$\underline{\hat{s}}(\underline{x})=\int_\theta \underline{s}(\underline{\theta})dP(\underline{\theta}/\underline{x}) \tag{4.6}$$

We obtain, similar to the previous results,

$$\ell(\underline{x}) = \frac{1}{c_1}\exp\{\int<d\underline{k},\ R\underline{\hat{s}}(\underline{x})>\} \tag{4.7}$$

and

$$dP(\underline{\theta}/\underline{x}) = c_1K(\theta)d\pi(\theta)\exp\{\int<d\underline{k},R(\underline{s}(\underline{\theta})-\underline{\hat{s}}(\underline{x}))>\}, \tag{4.8}$$

where we denote the inner product by <>. These results, are identical to those of Esposito for the Gaussian problem, [2]. They continue to hold for the multivariate analogues of the

Gamma, Poisson, and other members of the multivariate exponential family.

A weakness of the preceeding structure is that the estimate, say $\hat{s}_i(x)$, depends on all x_j, $j=1,2,\ldots i,i+1,\ldots N$. Consequently, if we want to make the association with time samples, $\hat{s}_i(x)$ is either unrealizable or can be estimated only after all the data is available. One way of avoiding this is to develop the expression for $\ell(\underline{x})$ recursively.

Let the superscript notation denote the past sequence, e.g., $x^n=\{x_1,x_2,\ldots x_n\}$. For purposes of notational simplicity we will also consider $\underline{s}(\underline{\theta})=\underline{\theta}$. We assume the following structure

$$\ell(x_n/x^{n-1},\theta^n) = \frac{p_1(x_n|x^{n-1},\theta^n)}{p_0(x_n|x^{n-1})} = K(\theta^n)g(x^n)e^{\theta_n f_1(x^n)} \quad (4.9)$$

and observe that

$$\ell(x_n/x^{n-1})=\int_{\theta^n} \ell(x_n|x^{n-1},\theta^n)dP(\theta^n)= \frac{p_1(x_n|x^n)}{p_0(x_n|x^n)} \quad (4.10)$$

In analogy to (2.2), the derivative with respect to x_n is

$$\frac{\partial\ell(x_n/x^n)}{\partial x_n} = \ell(x_n/x^{n-1})\left[\hat{\theta}_n \frac{\partial f_1}{\partial x_n} + \frac{1}{g(x^n)} \frac{\partial g}{\partial x_n}\right] \quad (4.11)$$

where $\hat{\theta}_n$ is now the causal MSE

$$\hat{\theta}_n(x) = E(\theta_n|x^n,H_1) = \int \theta_n dP(\theta_n|x^n) \quad (4.12)$$

The solution to (4.10) is

$$\ell(x_n|x^{n-1}) = \frac{1}{c_1} g(x^n) \exp\left\{\int \hat{\theta}_n \frac{\partial f_1}{\partial x_n} dx_n\right\} \quad (4.13)$$

For notational convenience, we define

$$F(x^n) \triangleq \int \hat{\theta}_n \frac{\partial f_1(x^n)}{\partial x_n} dx_n.$$

The conditional ALR becomes

$$\ell(x_n|x^{n-1})= \frac{1}{c_1} g(x^n)e^{F(x^n)} \quad (4.14)$$

The parameter c_1 no longer is a constant, it now depends on the previous data. To see this, note the definition of $\ell(x_n|x^{n-1})$ from (4.10), multiply through by $p_0(x_n|x^{n-1})$, and integrate with respect to x_n. We obtain

$$c_1=c_1(x^{n-1})=\int g(x^n)e^{F(x^n)}p_0(x_n/x^{n-1})dx_n \quad (4.15)$$

We are now in a position to develop the likelihood ratio recursively. We write

$$\ell(x^n) = \ell(x^n|x^{n-1})\ell(x^{n-1}) \quad (4.16)$$

Substitute (4.14) and take logarithms

$$L(x^n) = L(x^{n-1}) + \ell n\, g(x^n) + F(x^n) - \ell n\, c_1 \qquad (4.17)$$

This is the desired recursive formula which contains only causal estimates.

COMMENTS

The two examples given illustrate the relationships between the average likelihood ratio, the conditional mean estimate, and the a priori distribution. Clearly, a variety of examples can be constructed for the Poisson, beta, and binomial distributions, all of which are members of the exponential family. If the a priori distribution is unknown, the conditional mean estimate can not be specified. An engineering solution would then be to use another appropriate estimate (e.g., maximum likelihood) in the correlation operation. This is consistent with previous comments in the literature, but now applies to all cases where the conditional densities are members of the exponential family.

REFERENCES

1. Ferguson, T., Mathematical Statistics, Academic Press, 1967.

2. Esposito, R., "A Class of Estimators for Optimum Adaptive. Detection," Information and Control, Vol. 10, Feb. 1967, pp. 137-148.

   See also, "On a Relation Between Detection and Estimation in Decision Theory," Info. and Control, Vol. 12, Feb. 1968, pp. 116-120.

3. Balakrishnan, A.V., "On a Characterization of Processes for Which Optimal Mean-Square Estimates Are of Specified Form," IEEE Trans. on Info. Theory, Vol. IT-6, Sept. 1960, pp. 490-500.

CHARACTERIZATION OF NON-RENEWAL ERROR-PROCESSES

JEAN-PIERRE A. ADOUL,
University of Sherbrooke,
Sherbrooke, Que. Canada.

ABSTRACT

The paper is concerned with practical characterizations of digital communication channels for such use as predicting the performances of error control techniques. The attention is focused on the class of channel error processes in which interval length between errors cannot be assumed statistically independent, a commun situation with troposcatter channels. A statistic is presented which well describes the clustering properties. An approximation method is outlined for the determination of the probability of occurrence of any error pattern.

INTRODUCTION

In digital communication systems, much attention has been focused on minimizing transmission errors. In recent years, a wealth of error-correction techniques have been devised which hold the promise of achieving very low error rates with practical complexity requirements. The performance of a particular technique however, is heavily contingent upon the statistical properties of the error process of the channel in conjonction with which it is to be used. In fact, the technique should be tailored to these clustering properties if significant gain is to be obtained.

A meaningful characterization of error processes is often achieved with the use of renewal models, the occurrence of an error being the renewal event. In particular the one-error-state case of Fritchman's partitionned Markov-chain [1] has been used with success by several investigators to determine code performances. Renewal models are tractable and can very well account for the distribution of the length of error-free intervals henceforth referred as error gaps or simply gaps. The tail of this distribution (i.e.: the probability of very long error free runs) is a critical feature of the channel [2], for it determines which of the two detection-and-retransmission versus forward error correction is preferrable.

Inherent to all renewal models is the assumption of statistical independence between successive error gaps. This is due to the loss of the "model memory" at the time of an error.

As will be shown in the data analysis presented at the end of this paper, gap independence cannot be readily assumed for all channels. In particular, many troposcatter channels exhibit a substantial positive correlation between two gaps. Under such circumstances retaining a renewal model would be inadequate in many respects.

---

* Research supported by the Canadian Defense Research Board, Ottawa, Canada, under grants 2804-21, to University of Sherbrooke.

CHARACTERIZATION OF
NON-RENEWAL ERROR PROCESSES

MODELLING APPROACH

Several modelling approaches have been developed to deal with non-renewal channels [3][4]. In these approaches, channel characterization is achieved through a set of parameters associated with a model. The values of these parameters are optimized in order that some error distribution of the real channel may be best carried over by the model. Basically, model parameters have been optimized with respect to either the conditional gap distribution [4] (which is the distribution of the interval between error conditioned on the length of the previous interval), or the density (or counting) distribution [3]. The later distribution labeled P(m,n) is the probability of occurrence of m errors within n consecutive transmitted digits. In this paper, a statistic is presented which forms a unifying concept for model implementation. However, our main interest here lies more in the area of clustering characterization for practical purposes rather than in analytical modelling.

EMPIRICAL CODE EVALUATION

In addition to the analytic channel characterization approach just described, direct evaluation of code performance using measured error distributions is an approach often followed by those investigators who have become impatient with the slow development of analytic characterizations. Although it is a valid way of determining performance, it does not allow an understanding of the subtle effect that variations in the error patterns have on code evaluation. Moreover, each set of data represents only one point in a performance analysis, and it is seldom possible to extrapolate results. Most importantly, this procedure is not economical in that it is generally necessary to store large quantities of recorded data and to perform a large number of computations. For those who favor direct numerical evaluation of code performances on digital computer without reference to descriptive modelling this paper suggests an interesting intermediate approach using Multigap distributions. As will be shown, this approach is economical in that the most relevant properties of the real channel error patterns are summarized in a relatively small set of numbers. From these numbers almost any desired error distribution can be accurately generated, thereby making it possible to evaluate almost any code desired. In addition to economy, flexibility and accuracy, the number of computations required to evaluate a given code remains within reasonable bounds.

MULTIGAP: *Definition and Properties.*

Consider a stationnary binary error process

$$\{e_n | e_n = 0,1\} \text{ , where } e_n = 1$$

indicates the occurrence of a transmission error at time n. The interval between two consecutive errors, say $e_k = 1$ and $e_\ell = 1$ while $e_j \neq 1$ for $k < j < l$, defines a gap, G. The length of this gap is $\ell - k$. Note that an error sequence can be equivalently described in terms of its associated gap sequence $\{G_i | i \varepsilon I\}$, where subscript i indicates the order in which gaps occur, and where G_i, the length of the i^{th} gap, can assume any non-zero positive integer. Because of stationarity all gaps G_i, have the same distribution. The expected gap length (if it exists) can be shown, see ref. [2], to equal

$$E(G_i) = \frac{1}{Pr\{e_j = 1\}}$$

A renewal process (i.e.: independent gaps) is entirely defined by the gap-length distribution. Thus any feature of the channel can be inferred from this distribution. For instance the distribution of the total length of r consecutive gaps is readily obtained by performing r-1 convolutions of the gap length distribution with itself.

If however, the error process is non-renewal that is with dependent gaps, the distribution of the length of r consecutive gaps can deviate significantly from the result of r-1 convolutions. This deviation illustrates the way in which gaps associate for instance short gaps with short gaps, etc.

For this reason, the concept of gap is extended as follows. Let a multigap of order r be r consecutive gaps and let its length be the sum of the composite-gap length:

$$G^r = \sum_{k=i}^{i+r-1} G_i \; .$$

The multigap-length distribution, will be denoted

$$M(r,\ell) \triangleq Pr\{G^r = \ell\}$$

In particular, $M(r,\ell)$ will be set equal to zero unless ℓ is greater or equal to r.

Several analytical properties of multigap distributions are obtained in ref. [5]. In particular it is demonstrated that the probability, $P(m,n)$, of finding m errors in a block of length n can be expressed in terms of multigap distributions.

$$P(m,n) = \frac{1}{E(G)} \sum_{j=m-1}^{n-1} (n-1) \left[M(m-1,j) - 2M(m,j) + M(m+1,j)\right] \qquad (\dagger)$$

A converse relation is also shown to exist by which $M(m,n)$ can be expressed in terms of $P(m,n)$ distributions, thereby showing the formal equivalence between density and multigap characterization of binary processes.

The later characterization has the advantage of being more tractable and much easier to be derived from recorded error data. As a matter of fact, in the case of real channel error sequences we deal with one error in hundreds or thousands or perhaps millions of transmitted digits. Thus considering the gap sequence $\{G_n\}$ instead of the error sequence $\{e_i\}$ reduces, on the average, the number of random variables involved in an observed sequence by a large factor, $E(G)$. It follows that a compact manner of storing experimental data from error sources consists of recording in a file the length of successive gaps. Hence, tabulating the multigap frequencies is an easy "gap by gap" operation, while other statistics such as $P(m,n)$ require almost a "bit by bit" analysis of the binary sequence.

USE OF MULTIGAP STATISTICS

In this section, we analyze what information on the binary process is available in the multigap characterization and how it could be used to closely approximate the probability of occurrence of given error patterns, as these probabilities are needed in code evaluation.

An exhaustive characterization is the one that provides the probability

$$\Pr\{e_n = 1 \mid e_{n-1}, e_{n-2} \ldots\}$$

that the "present" random variable e_n be one knowing all past outcomes or equivalently the position of the last, second last... error.

Consider for instance the past sequence of Figure 1, the occurrence of a "1" at time n, would result in a gap of length $G^1 = 53$ and so on; such an event has probability.

$$\Pr\{G^1=g_1, G^2=g_2, G^3=g_3 \ldots\}=\Pr\{G^1=g_1\}.\Pr\{G^2=g_2 \mid G^1=g_1\}\ldots$$

It is clear that the latter probabilities cannot be expressed by the set of multigap distributions alone, indeed multigaps do not provide an exhaustive characterization.

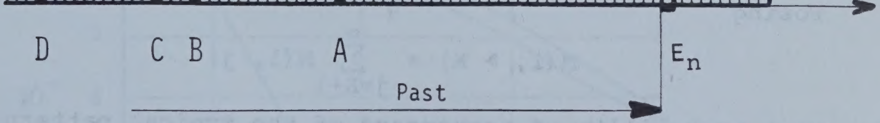


FIGURE 1 - Prediction from Multigaps

However, let us suppose that there are observers placed at each error in the past sequence: observer **A** at the last error, observer B at the second last, etc... Each of these observers can voice an opinion about the probability of digit e_n to be in error, based on the set of multigap distributions. In Figure 1, the probability of the event $e_n = 1$ is $M(1, 34)$ for observer A while it is $M(2, 49)$ for observer B, etc... Thus, the best we can do to estimate the probability of digit e_n being in error (from the multigap characterization) is a weighting of the observer's opinions. If only one of the observer's locations is known, of course, we will content ourselves with his opinion.

CODE EVALUATION

We now consider in what manner multigap characterization can be used for the specific problem of code evaluation. Computing the probability of error of a code used over a real channel, amounts to the evaluation of the frequency of occurrence of error patterns that cannot be corrected properly by the coding technique considered (uncorrectable patterns). Such patterns typically involve permissive number of errors in constrained lengths with conditions on the adjacent guard spaces, etc...

When the channel error process can be assumed to be renewal, the probability of such patterns can be obtained by convolutional method. Take for example the following pattern: m_1 errors in n_1 digits than m_2 errors in the next n_2 digits finally m_3 errors in the next n_3 digits $(m_i \leqslant n_i)$.

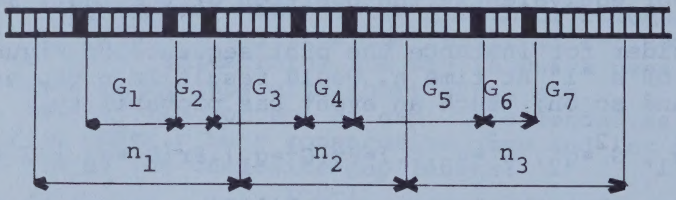


FIGURE 2 - Typical error pattern

The pattern starts at random, at i = 1 by convention. Integers n_1 , n_2 , n_3 , m_1 , m_2 , m_3 being fixed, it is possible to examine one typical error pattern, see Fig. 2, and compute its probability. It consists of a few "non-error" outcomes followed by a concatenation of $h = n_1 + n_2 + n_3 - 1$ gaps and again by a few "non-error" outcomes. Suppose the first error occurs at $i = k$ $(1 < k < n - m)$, k being a fixed distance from a random position, the later event has probability,

$$\Pr\{e_K = 1\} = \frac{1}{E(G)}$$

Posing

$$M(1, > K) = \sum_{j=K+1}^{\infty} M(1, j)$$

the exact probability of occurrence of the typical pattern, still when the renewal assumption can be made, would be:

$$P_T = M(1, > k-1)\left[\frac{1}{E(G)}\right] M(1,g_1).M(1,g_2)\ldots.M(1,g_h)M(1, > k_2)$$

The probability of the error pattern considered in our example would then be the sum of the probabilities of all possible such "typical" patterns,

$$P = \sum_{k=2}^{n_1-m_1+1} M(1, > k-1)\ \frac{1}{E(G)} \sum_{g_1=1}^{n_1-m_1-k+1} M(1, g_1) \sum\ \ldots\ldots.$$

Convolution calculus of this type can be readily performed on a digital computer with a few nested "Do loops" and is of practical interest.

Next, the question arises: can one evaluate the probability of error patterns in a similar way when the channel is non renewal and characterized by multigaps ? Here again we can decompose the typical pattern in a concatenation of events such as gaps and multigaps, and if one contents himself with reasonable approximations, the assumption of independence between these events would permit the use of convolution calculus. For instance the typical pattern of Figure 2 can be decomposed into 5 multigaps of respective order: n_1-1, 1, n_2-1, 1, n_3-1. Assuming the independence of these multigaps the probability of the typical pattern becomes:

$$P = M(1, > K-1) \frac{1}{E(G)} M(n_2-1, \ell_1) M(1, \ell_2) M(n_2-1, \ell_3)$$
$$M(1, \ell_4) M(n_3-1, \ell_4) M(1, >\ell_5)$$

We do not pretend that the above assumptions are right nor that this particular method of decomposition makes the best use of the multigap characterization. However, this approach offers a double advantage, that of retaining great simplicity, and that of keeping to a minimum the number of independence assumptions to be made, thereby improving the assumption of all gaps independent. In fact, oftentimes this approach turns out to be very accurate. Clearly, the fewer independence assumptions which are made, and the greater the order of the multigaps involved the more accurate is the result of this approximation method.

To illustrate the basic approach, P(m,n), the probability of m errors within n consecutive digits is calculated for a troposcatter channel. First (1), curve "P" on Figure 3 is the exact value of P(m,n) derived from the multigap distribution by means of Equation (†), which incidentally was verified to coincide perfectly with the direct measure of P(m,n) on the data.

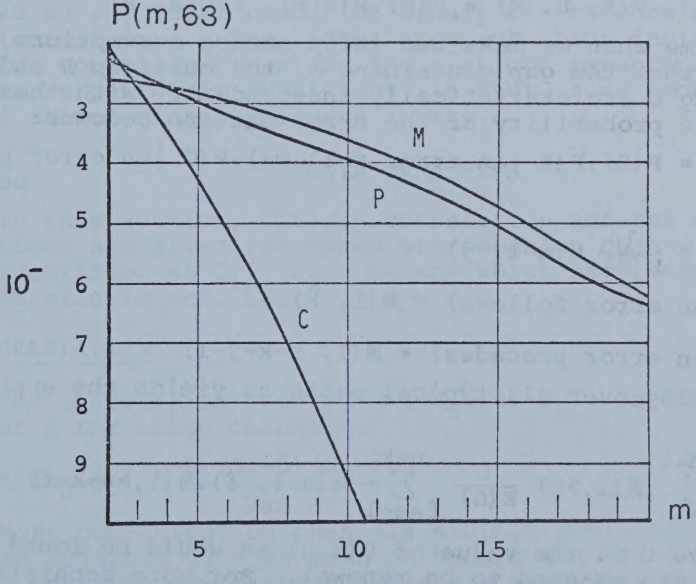


FIGURE 3 - P(m,n) Distribution.
P : exact distribution; A: approximation;
C : approximation assuming gap independency.

Secondly, an approximate value for $P(m,n)$ is calculated from the multigap distributions using the basic approach suggested above, curve A. Indeed, this calculation is of no practical interest since the exact value is obtainable from Equation (†)! We do this calculation however to allow observation of what is lost in the approximation. To begin, we consider a typical pattern of given m and n which would be like Figure 4. This error

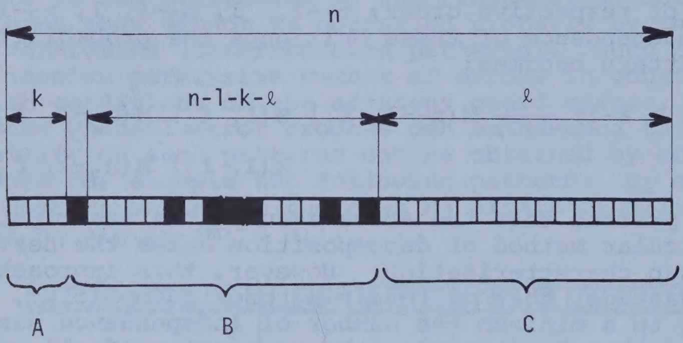


FIGURE 4 - Calculation of $P(m,n)$ by the approximate method.

pattern is decomposed in three concatenated events A, B, C.

A: k consecutive zeros.
B: an error followed by a multigap of order m-1 and length j.
C: n - k - j - 1 consecutive zeros.

The exact calculation would yield:

$$P(A, B, C) = P(B).P(A|B).\ P(C|A, B)$$

Now is the time when we make two independence assumptions, namely we assume that the gap containing A, the multigap B and the gap containing C are statistically independent. With these assumptions, the probability of the error pattern becomes:

$$P(A, B, C) = P(B).P(A\ |\text{an error follows}).P(C\ |\text{an error precedes})$$

where

$$P(B) = \frac{1}{E(G)} M(m-1, j)$$

$$P(A \text{ an error follows}) = M(1, k)$$

$$P(C \text{ an error precedes}) = M(1, n-k-j-1)$$

finally, summing over all typical patterns yields the approximation:

$$P(m,n) \simeq \sum_{k=1}^{n-m+1} K(1,>k)\ \frac{1}{E(G)} \sum_{\ell=m-1}^{n-k} M(m-1, \ell).M(1,>n-k-\ell)$$

Thirdly, curve C is the value of $P(m,n)$ as would be found if the channel were assumed to be renewal. For more sophisticated error patterns, the accuracy of the approximate method might not be as good as in the present case, however, the above comparison suggests some confidence in the method and exemplifies the fact that it can save one from colossal inaccuracies which could result from an illegitimate renewal assumption.

This investigation of evaluation of error-pattern probabilities will end with a few practical considerations concerning the use of digital computers in achieving the goal. First, it must be noticed that a good number of code-evaluation problems only require multigap lengths below, say, a hundred, and multigap orders up to 20 or so. In some special cases, however, it is worthwhile to implement blocks, or interleaving periods, of thousands of bits long. This is reasonable on channel with very low error rate where expected distances between error-bursts might be of about the same great length. Thus, by using very long blocks or interleaving periods one can make the best use of long error-free runs for "diluting" the effect of heavy bursts. Here again in many instances, multigaps of order up to 20 or 30 might turn out to be sufficient (occurrence of more than 20 or 30 errors in the block being exceedingly improbable). Moreover, and for the sake of economy, multigap probabilities of the same order and whose length lay between two large arbitrary values might be given the same averaged value. It follows that in many practical situations the necessary values of multigap distributions, $M(r, \ell)$, can be punched out on just 300 or 400 computer cards, a deck that summarized the channel clustering properties. Should an objective opinion be needed as to the reliability improvement some new code would cause if used on a given channel X, we can write a computer program to approximate and sum up the probabilities of the code incorrectable error-patterns and use the characterizing deck X as the data deck for this program. Finally, a simple comment is made as to what values of $M(r, \ell)$ should be punched out. A smoothed multigap distribution could be derived from the experimental data. However, a straight-forward alternate solution is to set $M(r, \ell)$ simply equal to the number of occurrences of the multigap, $G^r=\ell$, divided by the total number of binary digits considered in the experimental data. The latter approach which incidentally was used for computing $P(m,n)$ in the previous paragraph, reveals itself to be perfectly sufficient, in fact, some smoothing comes out of convolutional products anyway.

DATA ANALYSIS

In this section, the gap correlation and the multigap distributions are given for three troposcatter channels which have been identified as C_1, C_2, C_3 and which have been described in some details in ref. [2].

GAP CORRELATION

The correlation between Gap G_n and G_{n+r} is given as a function of r for three channels.

$$\mathrm{Corr}(G_o,G_r) = \frac{E[(G_o-\bar{G})(G_r-\bar{G})]}{\mathrm{Var}\ G}$$

Most communication channels exhibit positive correlation between adjacent gaps as seen on Figure 5-a and 5-b.

A positive correlation indicates the tendency of short gaps to be followed by short gaps while long gaps follow after long gaps. A negative correlation, on the contrary, would suggest

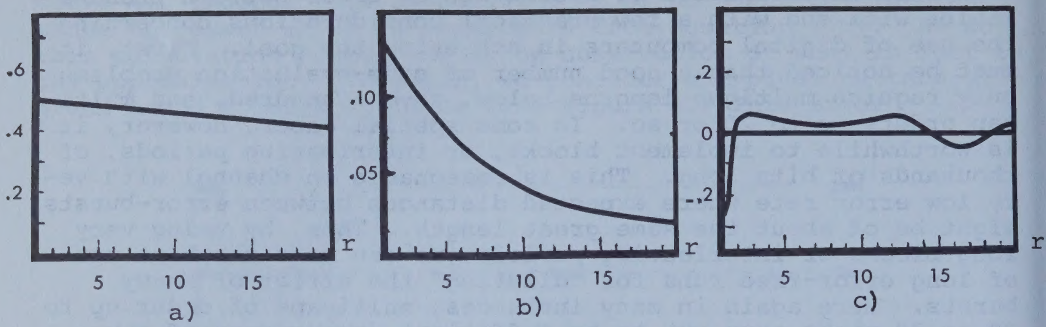


FIGURE 5 - Gap correlation

long gaps followed by short gaps and vice versa. Channel C_3, however, does not follow the general pattern. Here the strong negative correlation is due to the fact that for this high data rate channel of 10^6 bits/sec, an isolated error, or equivalently, two successive long gaps seldon occur. Still another way is to say that more than one error results at each occurrence of a physical cause of errors. Channel C_1 is highly non renewal as the gap correlation is still substantial accross hundreds of gaps. Channels C_1 and C_2 exhibit an exponential tail indicating that a Markov-chain type of gap dependence would be appropriate to these cases.

MULTIGAP DISTRIBUTION

The multigap (anticumulative) distribution

$$\Pr\{G^r > g\}$$

is given as a function of g for orders r = 1, 5, 10, 20.

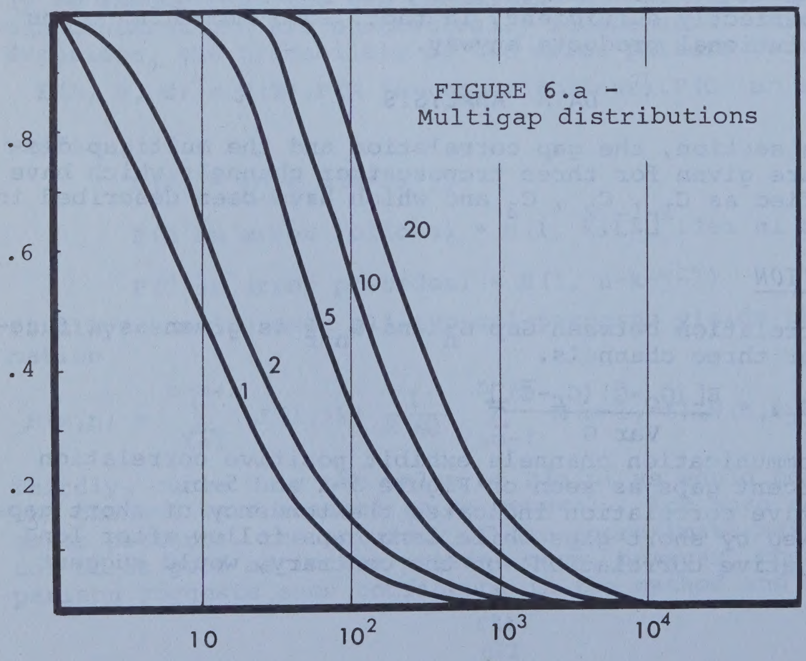


FIGURE 6.a - Multigap distributions

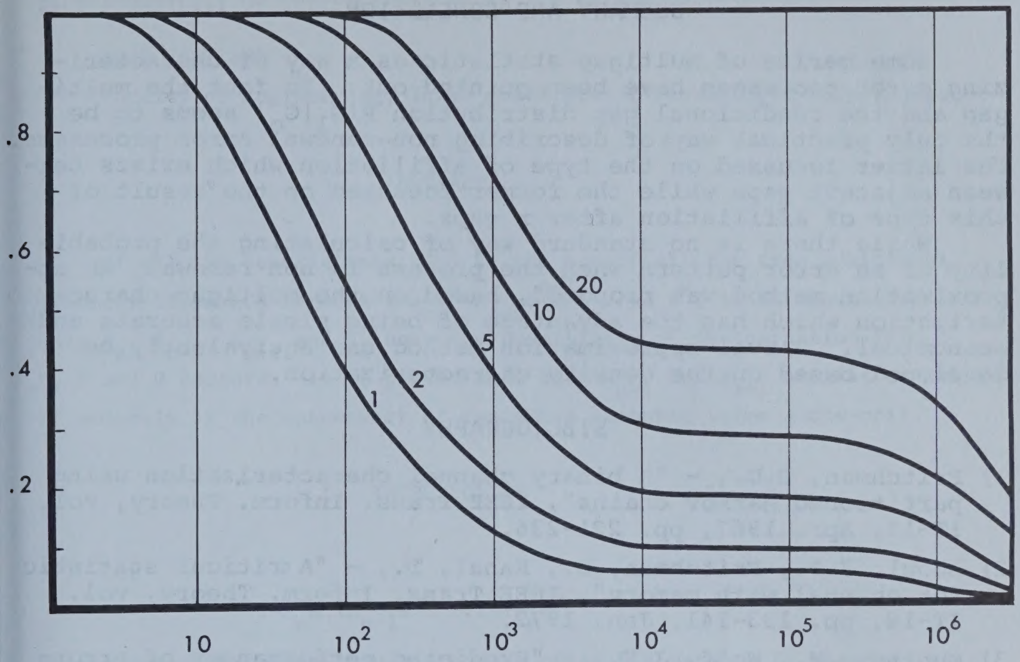


FIGURE 6.b - Multigap distributions - C_2

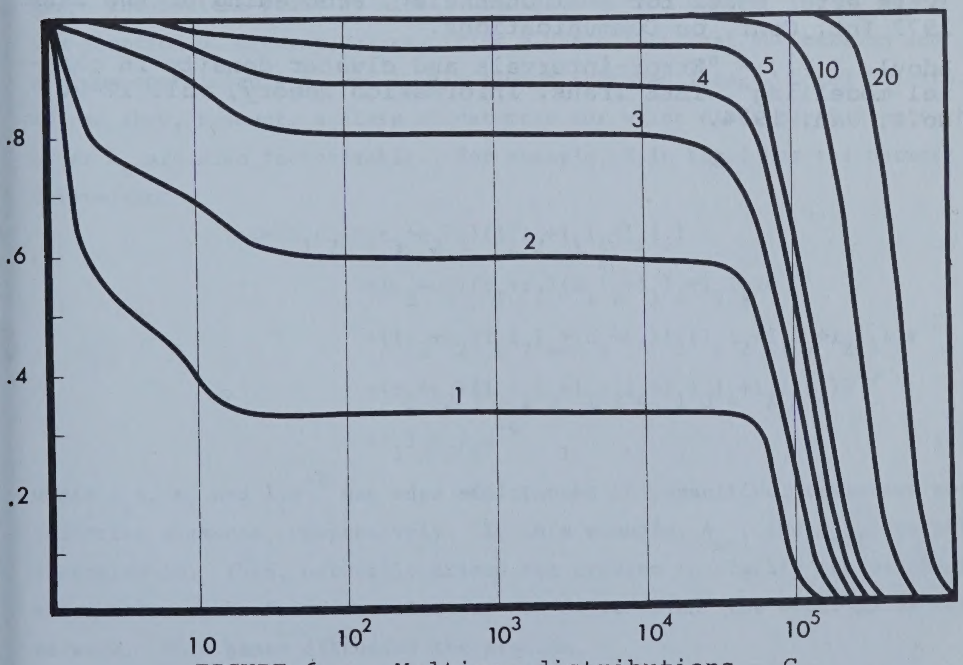


FIGURE 6.c - Multigap distributions - C_3

SUMMARY AND CONCLUSION

Some merits of multigap statistics as a way of characterizing error processes have been pointed out. In fact the multigap and the conditional gap distribution $P\{G_1|G_0\}$ seems to be the only practical ways of describing non-renewal error processes. The latter focussed on the type of affiliation which exists between adjacent gaps while the former focussed on the result of this type of affiliation after r gaps.

While there is no standard way of calculating the probability of an error pattern when the process is non renewal, an approximation method was proposed, based on the multigap characterization which has the advantage of being simple accurate and economical. A dual approximation method can equivalently be developed based on the density characterization.

BIBLIOGRAPHY

1) Fritchman, B.D., - "A binary channel characterization using partitioned Markov chains", IEEE Trans. Inform. Theory, vol. IT-13, Apr. 1967, pp. 221-236.

2) Adoul, J.P., Fritchman, B., Kanal, L., - "A critical statistic for channel with memory", IEEE Trans. Inform. Theory, vol. IT-18, pp. 133-141, Jan. 1972.

3) Muntner, M., Wolf, J.K., - "Predicted performances of error-control techniques over real channels", IEEE Trans. Inform. Theory, vol. 14, Sept. 1968, pp. 640-650.

4) Chien, R.T., Haddad, A.H., Coldbert, B., Moyes, E., - "An analytic error model for real channels", Proceeding of the IEEE 1972 Int. Conf. on Communications.

5) Adoul, J.P., - "Error-intervals and cluster density in channel modelling", IEEE Trans. Information Theory, vol. IT-20, No.1, Jan. 1974.

FACTORIZABILITY OF COEFFICIENTS OF AN LCR NETWORK DETERMINANT

YOJI KAJITANI

Faculty of Engineering, Tokyo Institute of Technology, Tokyo, Japan

ABSTRACT

Of all the coefficients $\{A_k\}$ in the sum of all the tree admittance products of an LCR network $T=A_n s^n + A_{n-1} s^{n-1} + \ldots + A_m s^m$, it is shown that only A_n, A_{n-1}, A_{m+1} and A_m have the possibility of factorization into L, C and R factors. It is proved that both A_{n-1} and A_{m+1} are factorizable if and only if the subnetwork of resistive elements forms a one-port.

I. INTRODUCTION

Let

$$T=A_n s^n + A_{n-1} s^{n-1} + \ldots + A_m s^m$$

be the sum of tree admittance products of an LCR network N, where s is the complex frequency. Because of its importance in network theory, there have been many investigations to this function, for example [1~6] from graph theoretical point of view. Considering each coefficient A_k to be the function of element values, it is a basic fact that the maximum and minimum coefficients A_n and A_m are always factored into L, C and R factors. We can show, however, a class of networks for which other coefficients than A_n or A_m are also factorizable. For example, N in Fig.1 has the network determinant

$$\begin{aligned} T=&(c_1c_2+c_1c_3+c_2c_3)(l_1l_2+l_1l_4+l_2l_4) \\ &+(c_2+c_3)(r_1+r_2)(l_1l_2+l_1l_4+l_2l_4)s^{-1} \\ &+\{(c_1+c_2)l_1l_2l_4+(c_1+c_3)l_3(l_1l_2+l_1l_4+l_2l_4)\}s^{-2} \\ &+(r_1+r_2)(l_1l_2l_3+l_1l_2l_4+l_1l_3l_4+l_2l_3l_4)s^{-3} \\ &+l_1l_2l_3l_4s^{-4} \end{aligned} \qquad (1)$$

where c_is, r_i and l_is^{-1} are edge admittances of capacitive, resistive and inductive elements, respectively. In this example, A_{n-1} and A_{m+1} are both factorizable. Then, naturally arises the problem to clarify the relationships between the factorizability of coefficients and the topology of the network. This paper discusses the problem.

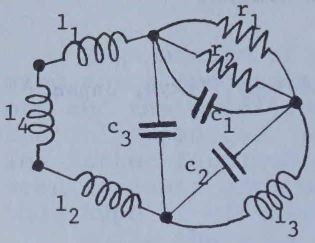


Fig.1 An LCR network for which A_{n-1} and A_{m+1} are totally factorizable

II. TOTALLY FACTORIZABLE COEFFICIENTS

Factorization of coefficients is defined as follows.

Def.1: Suppose X denotes C, R or L. A_k ($k=n,n-1,\dots,m$) is said to have X-factor if A_k is factored as

$$A_k=F^X\cdot F$$

where F^X and F are polynomials in element values, such that F^X is 1 or composed only of X-elelemnt values, and F does not contain any X-element values. Especially, if A_k has all C-, R- and L-factors as $A_k=F^C\cdot F^R\cdot F^L$, A_k is said to be totally factorizable.

For convenience, suppose N also denotes a graph corresponding to the network N. Each element value c_i, r_i and l_i are used to denote each corresponding edge. Let E_C, E_R and E_L be the sets of all the capacitive, resistive and inductive edges, respectively, and $E_N=E_C\cup E_R\cup E_L$. Suppose that each coefficient A_k also denotes the set of trees corresponding to the terms of $A_k s^k$. Let each tree be represented as the product of element values. Thus, in the above example, A_{-1} is regarded as a set of trees $\{c_2r_1l_1l_2, c_2r_1l_1l_4, \dots, c_3r_2l_2l_4\}$ of the graph N in Fig.1.

Let Q_t be the fundamental cut set matrix (mod 2) of N with respect to a tree t. t' is the cotree of t. Then Q_t can be written in the form

$$Q_t = \begin{array}{cccccc} t'\cap E_C & t'\cap E_R & t'\cap E_L & t\cap E_C & t\cap E_R & t\cap E_L \\ \left[Q^{CC} \right. & Q^{CR} & Q^{CL} & 1 & 0 & \left. 0 \right] \\ Q^{RC} & Q^{RR} & Q^{RL} & 0 & 1 & 0 \\ Q^{LC} & Q^{LR} & Q^{LL} & 0 & 0 & 1 \end{array}$$

To indicate a submatrix of Q_t, following convention is used: Let $E_1\subset t$ and $E_2\subset t'$. Then, $[E_1\times E_2]$ is a submatrix of Q_t that is the intersection of rows corresponding to the fundamental cut sets defined by the tree edges of E_1 and columns corresponding to the edges of E_2. For example,

$[(E_R \cap t)\times(E_L \cap t')]=Q^{RL}$.

It is well known that $t(E_2/E_1) \triangleq (t \cup E_2)-E_1$ is a tree if and only if the matrix $[E_1\times E_2]$ is nonsingular, or $\det[E_1\times E_2]=1 \pmod 2$.

Lemma 1. Suppose that A_k is totally factorizable. For any c, r, l $\in$ t and c', r', l' $\in$ t', the following equations hold in Q_t (t A_k).

$$[c\times r']\cdot[r\times c']=0 \tag{2}$$

$$[r\times l']\cdot[l\times r']=0 \tag{3}$$

$$[c\times l']\cdot[l\times c']=0 \tag{4}$$

$$\det[\{c,r,l\}\times\{c',r',l'\}]+[c\times c']\cdot[r\times r']\cdot[l\times l'] +[c\times r']\cdot[r\times l']\cdot[l\times c']+[c\times l']\cdot[r\times c']\cdot[l\times r']=0 \tag{5}$$

Proof: It is easy to show that $t(c'r'/cr) \in A_k$ if and only if $t(c'/c) \in A_k$ and $t(r'/r) \in A_k$. Hence

$$\det[\{c,r\}\times\{c',r'\}]=[c\times c']\cdot[r\times r']$$

By substituting this into the expanding formula

$$\det[\{c,r\}\times\{c',r'\}]=[c\times c']\cdot[r\times r']+[c\times r']\cdot[r\times c']$$

we obtain eq.(2). Eq.(3) and (4) are similarly obtained. Eq.(5) can be proved by substituting eq.(2), (3), (4) into the expanding formula of $\det[\{c,r,l\}\times\{c',r',l'\}]$.

Q.E.D.

This lemma leads the main theorem.

Theorem 1. A_k is not totally factorizable unless k=n, n-1, m+1 or m.

Proof: Let A_k be totally factorizable. In Q_t ($t \in A_k$), eq.(2) means that at least one of Q^{CR} and Q^{RC} must be zero matrix. And by eq.(3) and (4), zero must be one of Q^{RL} and Q^{LR}, and one of Q^{CL} or Q^{LC}. Now all the possible cases of $\bar{Q}_t$ such that $Q_t=[\bar{Q}_t\ 1]$ are shown in Fig.2.

Type A

$$\begin{array}{c|ccc} & c' & r' & l' \\ \hline c & * & * & * \\ r & 0 & * & * \\ l & 0 & 0 & * \end{array} \qquad \begin{array}{|c|c|c|} \hline * & 0 & 0 \\ \hline * & * & 0 \\ \hline * & * & * \\ \hline \end{array}$$

A_1 A_2

Type B

$$\begin{array}{|c|c|c|} \hline * & 0 & * \\ \hline * & * & 0 \\ \hline 0 & * & * \\ \hline \end{array} \qquad \begin{array}{|c|c|c|} \hline * & * & 0 \\ \hline 0 & * & * \\ \hline * & 0 & * \\ \hline \end{array}$$

B_1 B_2

Type C

$$\begin{array}{|c|c|c|} \hline * & 0 & 0 \\ \hline * & * & * \\ \hline * & 0 & * \\ \hline \end{array} \qquad \begin{array}{|c|c|c|} \hline * & * & 0 \\ \hline 0 & * & 0 \\ \hline * & * & * \\ \hline \end{array} \qquad \begin{array}{|c|c|c|} \hline * & 0 & * \\ \hline * & * & * \\ \hline 0 & 0 & * \\ \hline \end{array} \qquad \begin{array}{|c|c|c|} \hline * & * & * \\ \hline 0 & * & 0 \\ \hline 0 & * & * \\ \hline \end{array}$$

C_1 C_2 C_3 C_4

(* and 0 denote nonzero and zero matrices, respectively)

Fig.2 $\bar{Q}_t$ in the proof for Theorem 1

If $\bar{Q}_t$ is Type A, it is easily shown that k=n or m, completing the proof. Next, suppose that $\bar{Q}_t$ is B_1 of Type B. By substituting eq.(5) into the expanding formula of $\det[\{c,r,l\}\times\{c',r',l'\}]$, we obtain

$$[c\times l']\cdot[r\times c']\cdot[l\times r']=0$$

for any c, r, $l \in t$ and c', r', $l' \in t'$. This means that at least one of Q^{CL}, Q^{RC} and Q^{LR} must be zero matrix. Hence this case is reduced to Type A or C. Thus the theorem has only to be proved for Type C.

Consider C_1. If rank of $Q^{RL}=1$, it is easy to show that k=m+1, completing the proof. Hence we suppose that rank of $Q^{RL}\geqq 2$. In this case, we can show that Q^{RC} and Q^{LC} are both zero matrices. For, if $Q^{RC}\neq 0$, there exist some r_i, $r_j \in t$ and c', $l' \in t'$ such that

$$\det[\{r_i,r_j\}\times\{c',l'\}]=1.$$

This means that $t^*=t(c'l'/r_ir_j)\in A_k$. However, t and t* cannot simultaneously be included in A_k since A_k is totally factorizable. If $Q^{LC}\neq 0$, there exist r_i, r_j, $l_k \in t$ and c_l', l_m', $l_n' \in t'$ such that

$$\det[\{r_i,r_j,l_k\}\times\{c_l',l_m',l_n'\}]=1.$$

This also contradicts to A_k's total factorizability. Thus, Q^{RC}, Q^{LC} and Q^{LR} are all zero matrices. In this case, $\bar{Q}_t$ is reduced to A_1 of Type A. Hence k=n. For other cases of Type C, the theorem is similarly proved.

Q. E. D.

As a special case, following corollary can be proved.

Corollary. If N is composed of two kinds of elements, only A_n and A_m are totally factorizable.

III. CONDITIONS THAT A_{n-1} IS TOTALLY FACTORIZABLE

There are various expressions of the condition that A_{n-1} is totally factorizable. Before the theorem, some preliminaries are necessary.

Def.2: $N(E_a:E_b)$ (E_a, $E_b\subset E_N$ and $E_a\cap E_b=\phi$) is a graph that is obtained from N by contarcting all the edges in E_a and deleting those in E_b. $T(E_a:E_b)$ is the set of all the trees of the graph $N(E_a:E_b)$

Lemma 2. For a fundamental cut set matrix Q_t (t:arbitrary),

$$p_X \quad \text{rank of } N(\phi:E_X)-\text{rank of } N(E_X:\phi)$$
$$=\text{rank of } (Q^{XY}\ Q^{XZ})+\text{rank of } \begin{pmatrix}Q^{YX}\\ Q^{ZX}\end{pmatrix}$$

where X denotes C, R or L. Y and Z are C, R or L other than X such that $Y\neq Z$. Rank of a graph G is defined as the number of its tree edges.

The proof is easy and omitted.

For N in Fig.1, for example, $p_C=2$, $p_R=1$ and $p_L=2$. Notice that if $p_X=1$ the subnetwork composed of X-elements forms so called a one port.

Theorem 2. Following six conditions are all equivalent.

(a) A_{n-1} is totally factorizable.

(b) A_{m+1} is totally factorizable.

(c) In Q_t $(t \in A_n)$, rank of Q^{CR}+rank of Q^{RL}=1.

(d) In Q_t $(t \in A_m)$, rank of Q^{RC}+rank of Q^{LR}=1.

(e) $p_R=1$, or the resistive subnetwork forms a one port.

(f) $\max_{t \in T} (|t \cap E_R|) - \min_{t \in T} (|t \cap E_R|) = 1$.

A brief proof will be given. Suppose that (a) or (b) holds. By the further investigations (omitted here) of Q_t (t:arbitrary) appeared in the proof for Theorem 1, it is shown that

$$\text{rank of } (Q^{RC}\ Q^{RL}) + \text{rank of } \begin{pmatrix} Q^{CR} \\ Q^{LR} \end{pmatrix} = 1 \qquad (6)$$

if A_{n-1} or A_{m+1} is totally factorizable. Thus, by Lemma 2, we obtain (e). (c) follows from eq.(6) since Q^{RC} and Q^{LR} are both zero matrices in Q_t $(t \in A_n)$. Similarly, since Q^{CR} and Q^{RL} are both zero matrices in Q_t $(t \in A_m)$, (d) follows from eq.(6) either.

Next suppose that (e) holds. Put

$$T_0 = A_n s^n + A_{n-2} s^{n-2} + \ldots..$$

$$T_1 = A_{n-1} s^{n-1} + A_{n-3} s^{n-3} + \ldots..$$

Then, it is easily shown that

$$T_i = T(E_R:\phi)\ T(\phi:E_C \cup E_L)$$

$$T_j = T(\phi:E_R)\ T(E_C \cup E_L:\phi)$$

where (i,j)=(1,0) or (0,1). In any case, A_{n-1} has R-factor $T(\phi:E_C \cup E_L)$ or $T(E_C \cup E_L:\phi)$. Thus let $A_{n-1}=F^R \cdot F^{LC}$. Then, F^{LC} is the maximum degree coefficient of a reactance network $N(E_R:\phi)$ or $N(\phi:E_R)$. Hence A_{n-1} is totally factorizable. By the same reason, it is true that A_{m+1} is totally factorizable. Thus, (e) leads (a) and (b).

Next suppose (c) or (d). By Lemma 2, $p_R=1$. Hence (e).

It is obvious [1] that (f) is equivalent to (e). Thus, the proof is completed.

IV. CLASSIFICATION OF NETWORK DETERMINANTS

As we have seen, the possibility of total factorizability of the coefficients is closely related with the interconnection of C-, R- and L-subnetworks. Table 1 shows a classification of network determinants from this point of view. The validity of this table is easily proved.

	Totally factorizable coeff. $(A_n, A_{n-1}, \ldots, A_m)$	Interconnection of C, R, L (p_C, p_R, p_L)
I	(P) (n-m=0)	(0,0,0)
III	(P,P) (n-m=1)	(1,1,0) or (0,1,1)
III	(P,P,P) (n-m=2)	(1,1,1)
IV	(P,P,P,P) (n-m=3)	(2,1,1) or (1,1,2)
V	(P,P,I,...,I,P,P) and $n-m \geq 4$	(*,1,*) (*:arbitrary)
VI	(P,I,...,I,P) and $n-m \geq 2$	the remainings

P (or I) means that totally factorization of the corresponding coefficient is possible (impossible)

Table 1. A classification of LCR network determinants

V. DETERMINATION WHETHER (A_k, A_{k-1}) CONTAINS A_n OR A_m

Consider a rather artificial* problem to decide by the factorization whether we can say a given successive coefficients (A_k, A_{k-1}) contains A_n or A_m or neither.

With respect to their total factorizability, all possible cases are classified into

(A) $(A_k, A_{k-1}) = (I, I)$.

(B)) $(A_k, A_{k-1}) = (P, I)$ or (I, P).

(C) $(A_k, A_{k-1}) = (P, P)$.

In case of (A), it can be say that neither A_n nor A_m is contained in (A_k, A_{k-1}). In case of (B) or (C), it is unable from Table 1 to determine whether (A_k, A_{k-1}) contains A_n or A_m or neither. However, if we investigate their factors in more detail, the problem can be solved as in the following theorems.

Theorem 3. Suppose that $(A_k, A_{k-1}) = (P, I)$ (or $= (I, P)$). Then, $(A_k, A_{k-1}) = (A_{n-1}, A_{n-2})$ $(= (A_{m+2}, A_{m+1}))$ of Type V in Table 1 if and only if A_{k-1} (A_k) has the R-factor.

Theorem 4. Suppose that $(A_k, A_{k-1}) = (P, P)$. Then, $(A_k, A_{k-1}) = (A_{n-1}, A_{m+1})$ if and only if neither C-factors nor L-factors of A_k and A_{k-1} are in common.

The proofs for these theorems are omitted.

---

* There is a theorem[6]: (A_k, A_{k-1}) determines the graph N within 2-isomorphisms if and only if (A_k, A_{k-1}) contains neither A_n nor A_m.

The problem left unsolved is to determine which of A_n and A_m is contained in (A_k,A_{k-1}) after we know that (A_k,A_{k-1}) includes A_n or A_m by the way given in the above discussion. As is easily seen, the difficulty exists when $(A_k,A_{k-1})=(P,P)$ of III, IV or V, since in these cases both $(A_k,A_{k-1})=(A_n,A_{n-1})$ and $(A_k,A_{k-1})=(A_{m+1},A_m)$ may occur.

The answer is negative. To show the situation, we have only to provide an example which shows that there exist two networks for which A_n and A_{n-1} of the one network are respectively the same as A_{m+1} and A_m of the other. The network N in FIg.3 has the network determinant T as

$$\begin{aligned} T=&c_1c_2c_3(r_1+r_2)s^3+c_1c_2c_3(l_1+l_2)s^2 \\ &+(r_1+r_2)\{(c_1c_2+c_1c_3+c_2c_3)l_4+(c_1c_2+c_1c_3)l_2+c_2c_3l_1\}s \\ &+(c_1c_2+c_1c_3+c_2c_3)(l_1l_2+l_1l_4+l_2l_4) \\ &+(c_2+c_3)(r_1+r_2)(l_1l_2+l_1l_4+l_2l_4)s^{-1} \end{aligned} \tag{7}$$

We see that A_0 and A_{-1} in the above are respectively equal to A_0 and A_{-1} in eq.(1) for N in Fig.1. Thus, we cannot in general say which of A_n and A_m is included by any means.

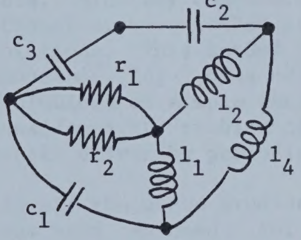


Fig.3 An LCR network for which $(A_3,A_2,A_1,A_0,A_{-1})=(P,P,I,P,P)$

VI. CONCLUDING REMARKS

Relationships between the factorizability of coefficients of the network determinant of an LCR network and the distribution of C-, R- and L-elements in the network are studied. The results include the theorems that a coefficient A_k cannot be totally factorized unless k=n, n-1, m+1 or m and that the total factorizability of A_{n-1} is the necessary and sufficient condition that the network is a reactance network terminated with a resistive one port.

REFERENCES

[1] S. L. Hakimi and W. Mayeda,"On coefficients of polynomials in network

functions," IRE Trans. Circuit Theory, vol.CT-7, Mar. 1960, pp.40-44.

[2] S. L. Hakimi,"Graphs with two kinds of elements," J. Franklin Inst., vol.271, 1960, pp.451-467.

[3] S. L. Hakimi,"On therealizability of set of trees," IRE Trans. Circuit Theory, vol.CT-8, Mar. 1961, pp.11-17.

[4] G. Kishi and Y. Kajitani,"On realization of tree graphs," IEEE Trans. Circuit Theory(Corresp.), vol.CT-15, Sept. 1968, pp.271-273.

[5] Y. Kajitani,"Graph theoretical properties of the node determinant of an LCR network,"IEEE Trans. Circuit Theory, vol.CT-18, May 1971, pp. 343-350.

[6] Y. Kajitani and Y. Hirano,"On realization of an LCR network from coefficients of a network determinant," (in Japanese) to be published in IECE of Japan.

EFFICIENT ANALYSIS OF NONLINEAR NETWORKS BY PIECEWISE-LINAR APPROXIMATION

I. HAJJ, K. SINGHAL AND J. VLACH
Department of Electrical Engineering
University of Waterloo, Ontario, Canada

ABSTRACT

A method is given for the efficient computation of time responses of nonlinear networks. The method depends on approximating the characteristics of the nonlinear elements by piecewise-linear curves. The analysis of a nonlinear network is thus replaced by the analysis of a sequence of linear ones. An efficient numerical Laplace inversion method with a variable-order variable-step algorithm is used to compute the response of the network in the operative linear modes and an iterative Newton-Raphson algorithm is used to detect the time when the network changes from one linear mode to another.

I. INTRODUCTION

The study of piecewise-linear (PWL) networks has attracted the interest of many network theorists in recent years. Most of the work has been carried out in the area of resistive networks, and many useful results have been obtained regarding the existence and the numerical computation of solutions [1-8]. The study of PWL dynamic networks, however, has not received as much attention and few significant results on these networks are available. Minorsky considered PWL oscillators [9]; he stated that the point-transformation method can be used to check for the existence of periodic solutions. This method has been studied in more detail by Vojtášek and Janáč [10]. Chua used PWL approximation to compute the transient solutions of simple nonlinear circuits [11]. More recently Armanazi considered the problem of finding the steady-state solutions of PWL networks driven by periodic inputs [12].

In all the above studies the state formulation of the network equations has been assumed. For PWL networks these equations are of the form

$$\dot{\underset{\sim}{x}} = \underset{\sim}{A}_i \underset{\sim}{x} + \underset{\sim}{B}_i \underset{\sim}{u} + \underset{\sim}{C}_i \underset{\sim}{u}_i \qquad i = 1,2,..,r \qquad (1)$$

where $\underset{\sim}{x}$,$\underset{\sim}{u}$, and $\underset{\sim}{u}_i$ are, respectively n,m, and ℓ vectors, $\underset{\sim}{A}_i$, $\underset{\sim}{B}_i$ and $\underset{\sim}{C}_i$ are, respectively n x n, n x m, and n x ℓ matrices. $\underset{\sim}{A}_i$, $\underset{\sim}{B}_i$, $\underset{\sim}{C}_i$ and $\underset{\sim}{u}_i$ are constant within region i. $\underset{\sim}{u}$ is the vector of applied inputs and $\underset{\sim}{u}_i$ is a vector generated by the element characteristics in region i. r is the total number of regions.

In this paper we shall describe a method of obtaining the transient responses of PWL networks. The method is not restricted to the state-space formulation. The fact that the network equations are linear within a given region will be exploited, and as a result, an efficient numerical Laplace transform inversion technique is used to find the response within the region. The basic idea used is to find the trajectory followed by the network variables in a given region in which the network is operating, solve for the time when the trajectory crosses a boundary, update the network element values to correspond to the region entered, and then find the trajectory followed in that region.

II. PIECEWISE-LINEAR ANALYSIS

A PWL network is defined as a network in which all the elements have PWL characteristics. Independent sources in such a network, however, may be characterized by arbitrary functions of time. In this paper we consider the PWL elements to be time-invariant. Thus a two-terminal PWL resistor (capacitor, inductor) is characterized by a PWL graph with finite linear segments in the i-v (q-v, ϕ-i) plane (see Table I, column 2). In analyzing PWL networks one is mainly interested in studying two problems. One is finding the equilibrium or the dc solutions; the other is finding the dynamic behaviour of the network over a finite interval of time. In the following, we review first some of the dc analysis methods used and then present the proposed method of transient analysis.

A. DC Analysis

DC solutions of a network are computed by solving a resistive network derived from the original one by open-circuiting the capacitors and short-circuiting the inductors. The equations that describe a PWL resistive network can be expressed in general as

$$\underset{\sim}{M}_i \underset{\sim}{x} + \underset{\sim}{w}_i = \underset{\sim}{u} \qquad i = 1,2,\dots,r \qquad (2)$$

where $\underset{\sim}{x}$ is a real n-vector which represents the network variables, and $\underset{\sim}{u}$ is the input vector. The real Euclidian space $\tilde{R}^n$ is divided into a finite number of regions by a finite number of boundary hyperplanes. In Eq. (2), r is the total number of the finite regions. $\underset{\sim}{M}_i$ is a constant nxn matrix, and $\underset{\sim}{w}_i$ is a constant n-vector. Both $\underset{\sim}{M}_i$ and $\underset{\sim}{w}_i$ characterise the network in region i. For example, if one uses the tableau approach in formulating the network equations in a given region, one obtains [13]:

$$\text{(KVL)} \qquad \underset{\sim}{A}^t \underset{\sim}{v}_n - \underset{\sim}{v}_b = \underset{\sim}{e} \qquad (3,a)$$

$$\text{(KCL)} \qquad \underset{\sim}{A}\, \underset{\sim}{i}_b = \underset{\sim}{j} \qquad (3,b)$$

$$\text{(Branch relations)} \qquad \underset{\sim}{P}_i \underset{\sim}{v}_b + \underset{\sim}{Q}_i \underset{\sim}{i}_b + \underset{\sim}{c}_i = \underset{\sim}{o} \qquad (3,c)$$

where Eqs. (3,a) and(3,b) are topological equations, $\underset{\sim}{A}$ is the incidence matrix, $\underset{\sim}{A}^t$ is the transpose of $\underset{\sim}{A}$, $\underset{\sim}{v}_n$ is the vector of node voltages, $\underset{\sim}{v}_b$ and $\underset{\sim}{i}_b$ are, respectively, the branch voltages and currents. Eq. (3,c) is a linear equation which characterizes the network in region i. Eq. (3) can be rewritten in the form

$$\begin{bmatrix} \underset{\sim}{A}^t & -\underset{\sim}{I} & o \\ o & o & \underset{\sim}{A} \\ o & \underset{\sim}{P}_i & \underset{\sim}{Q}_i \end{bmatrix} \begin{bmatrix} \underset{\sim}{v}_n \\ \underset{\sim}{v}_b \\ \underset{\sim}{i}_b \end{bmatrix} + \begin{bmatrix} \underset{\sim}{o} \\ \underset{\sim}{o} \\ \underset{\sim}{c}_i \end{bmatrix} = \begin{bmatrix} \underset{\sim}{e} \\ \underset{\sim}{j} \\ \underset{\sim}{o} \end{bmatrix} \qquad (4)$$

which is of form (2). The boundary hyperplanes dividing the regions are determined by the breakpoints on the characteristic of each PWL element.

Eq. (2) is usually solved by means of a modified Newton-Raphson (N-R) method [1-4]. Starting with an initial guess $\underset{\sim}{x}_k$ in region i, one solves (2) to get

$$\underset{\sim}{x}_{k+1} - \underset{\sim}{x}_k = \lambda \; [\underset{\sim}{M}]^{-1} \; (\underset{\sim}{u} - \underset{\sim}{u}_k) \tag{5}$$

where $$\underset{\sim}{u}_k = [\underset{\sim}{M}]_i \, \underset{\sim}{x}_k + \underset{\sim}{w}_i \tag{6}$$

and where λ is a scalar. If $\lambda=1$, then the unmodified N-R iteration formula is obtained [2]. If λ is chosen so as to minimize, or reduce, the quantity $||\underset{\sim}{u} - \underset{\sim}{u}_{k+1}||$, a modified N-R formula results [3]. If λ is chosen so that $\underset{\sim}{x}_{k+1}$ is on the boundary of region i, then one gets Katzenelson's algorithm [1,4]. The convergence of Katzenelson's algorithm has been studied in detail [6,7]. It has been extended to construct input-output characteristics and to analyze networks with multiple solutions [4]. It has also been applied to analyze large networks using sparse matrix solution methods with dyad updating of the LU factors [7,14].

B. Transient Analysis

The problem of finding the transient solution can be stated as follows: Given a network with known initial conditions at time t=o (these initial conditions may be obtained by finding the dc solutions at t=o), find the response of the network (voltages and currents) at some sampled points of time within a time interval T. The basis of the solution method that we shall adopt for analysing PWL networks is as follows: starting at t_o=o with the network operating in a known initial region, say i, and with given initial conditions, find the response of the network at the next point $t_1=t_o+h$, where h is a given time step. If the solution at t_1 is within region i, the initial conditions are updated to correspond to the values at t_1 and the response is computed at $t_2 = t_1 + h$. If the solution falls outside region i, the approximate time when the trajectory crosses the boundary is found, and thus the time step h is adjusted. The network element values and initial conditions are now updated to correspond to the new region and the network is again solved.

Since the network equations are linear within a region, linear techniques can be used to find the response within it. In our approach we use an efficient technique for the numerical inversion of Laplace transforms [15,16,17]. This technique has been extended to produce a variable-step size variable-order algorithm [18]. The inversion technique is equivalent to a few chosen frequency analyses of the network, and the special attention should be given to representing linear and PWL elements with initial conditions in the frequency domain. This representation is given in Table 1. Within each linear region the Tableau equation (3) now has the form

$$\begin{bmatrix} \underset{\sim}{A}^t & -\underset{\sim}{I} & 0 \\ 0 & 0 & \underset{\sim}{A} \\ 0 & \underset{\sim}{P}_i(s) & \underset{\sim}{Q}_i(s) \end{bmatrix} \begin{bmatrix} \underset{\sim}{V}_n(s) \\ \underset{\sim}{V}_b(s) \\ \underset{\sim}{I}_b(s) \end{bmatrix} = \begin{bmatrix} \underset{\sim}{E}(s) \\ \underset{\sim}{J}(s) \\ \underset{\sim}{C}_i(s) \end{bmatrix} \tag{7}$$

We now briefly outline the method.

Let f(t) be a time domain function and F(s) its Laplace transform, then

$$f(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(s)\, e^{st} ds = \frac{1}{2\pi jt} \int_{c'-j\infty}^{c'+j\infty} F(z/t) e^{z} dz \tag{8}$$

Using a Padé approximation to e^z with a numerator of degree N and a denominator of degree M, the approximate time domain function is given as

$$\hat{f}(t) = -\frac{1}{t}\sum_{i=1}^{M} k_i \; F(z_i/t) \tag{9}$$

where z_i are the poles of the Padé approximation and k_i the corresponding residues. This formula is valid when $n+N+2 \leq m+M$ where n,m are the number of finite seros and poles of F(s). For network problems this condition is met when $N+2 \leq M$.

It has been shown [17] that the method exactly inverts the first M+N+1 terms of the Taylor expansion of f(t). In addition if $N+2 \leq M$ the impulse function and some of its derivatives can also be inverted, ie., f(t) can have the form $f(t) = a\delta(t) + g(t)$. This property will be used later to create a NR scheme for finding derivatives of function.

The derivative of the function f(t) has an image in the transform plane given by

$$L\,[f'(t)] = sF(s) - f(o+)$$

and the second term to the right generates an impulse at t=o. As shown in [17] if an appropriate N,M are selected the approximate inversion relation becomes

$$\hat{f}'(t) = -\frac{1}{t^2}\sum_{i=1}^{M} k_i \, z_i \, F(z_i/t) \tag{10}$$

and the derivative is obtained at almost no additional computational cost as most of the effort involved is in finding $F(z_i/t)$.

In the context of the solution of linear and PWL networks the time domain solution within a linear region is equivalent to solving the Tableau equations (5) at a 'few' (usually from 1 to 24 depending on accuracy requirements) pretabulated complex frequency points and performing the sum indicated in (9).

III. BOUNDARY INTERSECTION

One of the problems that arises in the above method is the question of finding the intersection of the time response trajectory with the region boundaries. Some remarks are now in order: (1) By using the inverse Laplace transform method, the response in a given region is not obtained as a function of time, but rather is computed at specified time points, hence some iterative technique is needed in order to find the intersection with the boundary; (2) since the response trajectory may intersect the boundary more than once, i.e., it may leave and then reenter the region, the first time the trajectory intersects the boundary should be recorded.

In order to find the intersection point of the time response with the boundary, we apply a N-R algorithm together with (9) and (10) and the equation of the boundary. In practice, however, this iterative method may slow down the process of finding the transient solution, especially if many PWL elements are included in the network. Alternately, an approximate intersection point may suffice. A trade-off between speed and the error generated due to the inaccuracy of the intersection point should be taken into consideration.

The time step control is checked in two ways. The information derived from the linearized network within a region and specified allowable error in the numerical inversion is used in one case, while in the second the step size is obtained by finding the cross-over point on the PWL characteristics.

IV. COMPUTER PROGRAM IMPLEMENTATION

The above method has been implemented as a part of a general purpose computer program, WATAND, [9]. The attractiveness of the method is that a unified formulation and solution technique for both linear and non-linear networks is used. The program uses sparse matrix solution techniques and is capable of performing frequency domain analysis as well as both small-signal and large-signal analyses in the time-domain. A number of problems have been analyzed using the above computation method. These problems include networks with general nonlinear elements and transistor networks.

Conclusions

A method has been presented for computing the time responses of PWL networks. The method relies on the fact that the network is linear within the regions of the PWL characteristics. An efficient numerical Laplace inversion method with variable-order variable-step algorithm is used to compute the response in the linear region, and an NR iterative method is used to find the cross-over point at the boundaries of the PWL characteristics. Further possible applications of the method might be used in studies on distortion, noise, sensitivity, steady-state periodic solutions and oscillator problems.

REFERENCES

[1] J. Katzenelson, "An Algorithm for Solving Nonlinear Resistive Network", Bell Sys. Tech. J. Vol. 44, Oct. 1965, pp.1605-1620.

[2] L. Chua, "Efficient Computer Algorithm for Piecewise-Linear Analysis of Resistive Nonlinear Networks", IEEE Trans. on Circuit Theory, Vol. CT-18, No.1, Jan. 1971, pp.73-85.

[3] T. Ohtsuki and N. Yoshida, "DC Analysis of Nonlinear Networks Based on Generalized Piecewise-linear Characterization", IEEE Trans. on Circuit Theory, Vol. CT-18, No.1, Jan.1971, pp. 146-152.

[4] E.S. Kuh and I.N. Hajj, "Nonlinear Circuit Theory: Resistive Networks", Proc. IEEE, Vol. 59, March 1971, pp. 340-355.

[5] T. Fujisawa and E.S. Kuh, "Some Results on Existence and Uniqueness of Solutions of Nonlinear Networks", IEEE Trans. on Circuit Theory, Vol. CT-18, No. 5, Sept. 1971, pp. 501-506.

[6] T. Fujisawa and E.S. Kuh, "Piecewise Linear Theory of Nonlinear Networks", SIAM J. App. Math, Vol.22, No.2, March 1972, pp. 307-328.

[7] T. Fujisawa, E.S. Kuh, and T. Ohtsuki, "A sparse matrix method for analysis of piecewise-linear resistive networks", IEEE Trans. on Circuit Theory, Vol. CT-19, No.6, November 1972, pp. 571-584.

[8] I.N. Hajj, D.J. Roulston and P.R. Bryant, "DC Analysis of Transistor Circuits Via a Three-Terminal Piecewise-Linear Modelling Approach", International Symp. on Circuit Theory, Toronto, Ont., Canada, April 1973.

[9] N. Minorsky, "Theory of Nonlinear Control Systems", New York: McGraw Hill, 1969.

[10] S. Vojtášek and K. Janáč, "Solution of Nonlinear Systems", Iliffe Books Ltd., London, England, Academia, Prague, 1969.

[11] L.O. Chua, "Introduction to Nonlinear Network Theory", New York: McGraw Hill, 1969.

[12] A.N. Armanazi, "Steady-state Analysis of Piecewise-Linear Systems with Periodic Inputs", Proc. IEEE, Vol.61, No.6, June 1973, pp. 789-790.

[13] G.D. Hachtel, R.K. Brayton and F.G. Gustavson, "The Sparse Tableau Approach to Network Analysis and Design", IEEE Trans. on Circuit Theory, Vol. CT-18, No.1, January 1971, pp.101-113.

[14] I.N. Hajj, "Updating Method for LU Factorisation", Electronics Letters, Vol. 8, No.7, April 1972, pp. 186-188.

[15] J. Vlach, "Numerical method for transient responses of linear networks with lumped, distributed or mixed parameters", J. Franklin Institute, Vol. 288, pp. 99-113, 1969.

[16] K. Singhal and J. Vlach, "Program for Numerical Inversion of Laplace Transform", Electronics Letters, Vol. 7, pp. 413-415, 1971.

[17] M.R. Nakhla, K. Singhal and J. Vlach, "Simple method for numerical inversion of Laplace transform", Proc. Sixteenth Midwest Symposium on Circuit Theory, pp. XIV.5.1-9, April 1973.

[18] K. Singhal and J. Vlach, "Computation of Time domain response by numerical inversion of the Laplace transform". To be published.

[19] I. Hajj, K. Singhal, J. Vlach and P.R. Bryant, "WATAND - A program for the analysis and design of linear and piecewise-linear networks", Midwest Symposium on Circuit Theory, Waterloo, Ontario, Canada, April, 1973.

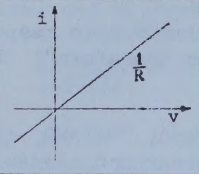
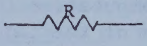
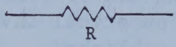
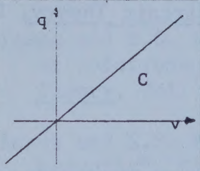
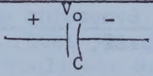
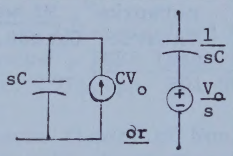
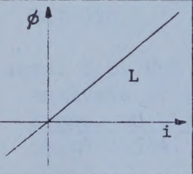
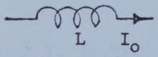
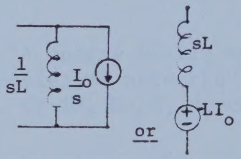
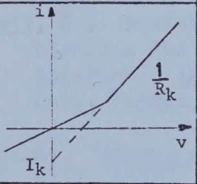
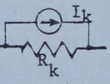
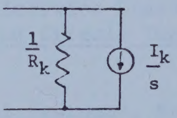
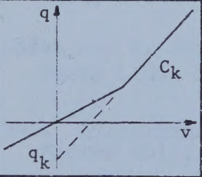
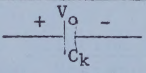
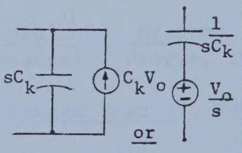
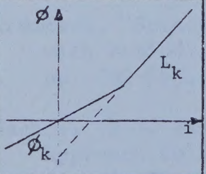
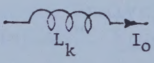
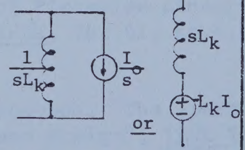
Element	Characteristics	Time Domain Representation	Frequency Domain Representation
Resistor		 v=Ri	
Capacitor		 $q = Cv$ $i = \frac{dq}{dt} = C\frac{dv}{dt}$	 sC CV_o $\frac{1}{sC}$ $\frac{V_o}{s}$ or
Inductor		 $\phi = Li$ $v = \frac{d\phi}{dt} = L\frac{di}{dt}$	 $\frac{1}{sL}$ $\frac{I_o}{s}$ sL LI_o or
PWL Resistor		 $i = I_k + \frac{1}{R_k}v$ (in region k)	 $\frac{1}{R_k}$ $\frac{I_k}{s}$
PWL Capacitor		 $q = q_k + C_kv$ $i = \frac{dq}{dt} = C_k\frac{dv}{dt}$ (in region k)	 sC_k C_kV_o $\frac{1}{sC_k}$ $\frac{V_o}{s}$ or
PWL Inductor		 $\phi = \phi_k + L_ki$ $v = \frac{d\phi}{dt} = L_k\frac{di}{dt}$ (in region k)	 $\frac{1}{sL_k}$ $\frac{I_o}{s}$ sL_k L_kI_o or

TABLE I

Representation of linear and PWL elements in the time and frequency domains

EFFECT OF AUGMENTING REACTIVE ELEMENTS IN AN IMPROPER NETWORK ON ITS STATE EQUATIONS

MOIEZ A. TAPIA
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT

The effect of augmenting the reactive elements in an improper network on the entries of the A and B matrices in its state equations is investigated. If the state equations for a network are already obtained, then the state equations for an augmented network can be easily and quickly obtained using the Tables given in the paper.

INTRODUCTION

The effect of modifying the reactive elements in a proper network, on the already found state equations was studied earlier by Moad [1]. The paper investigates the same effect for linear, time-invariant, improper network. The augmentation of essential as well as excess reactive elements is considered. The changes in the A and B matrices of the already found state equations due to modification of reactive elements are given in the Tables that follow later.

It will be assumed that the capacitors (or inductors) considered for augmentation belong to a capacitor loop (or inductor cutset).

Notation:

Let C_1, C_2, ---, C_{n_c} be n_c capacitors in a capacitor loop with voltage sources, if any, in it shortcircuited.

C_T denotes sum of all tree admittance products of the loop.

$C_{T, i1, i2, ---, ik}$, $1 \leq k \leq n_c$, denotes sum of tree admittance products of the loop with each tree containing capacitances C_{i1}, C_{i2}, ---, C_{ik}.

C_p denotes the product of all the capacitances in the loop.

$C_{p, i1, i2, ---, ik}$, $1 \leq k \leq n_c$, denotes the product of capacitances C_{i1}, C_{i2}, ---, C_{ik} in the loop. Observe that

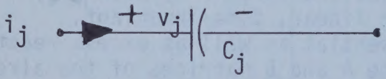
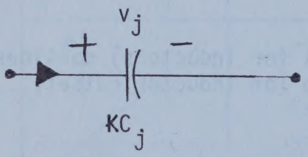
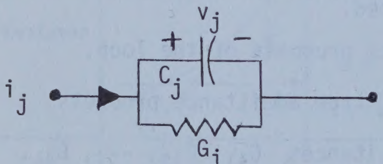
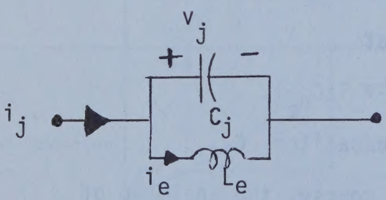
$$C_T = C_p\ (1/C_1 + 1/C_2 + ---+ 1/C_{n_c}).$$

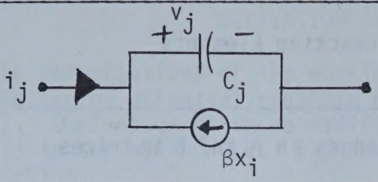
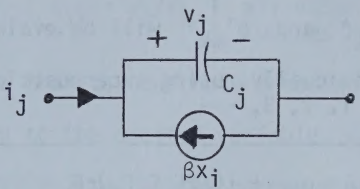
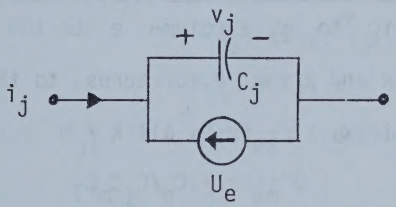
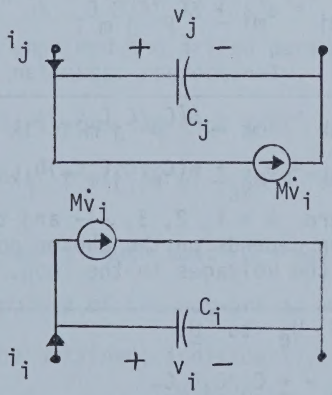
C_{Lj} denotes a capacitor loop that includes capacitor C_j.

L_T, L_p, etc. can be defined analogously. Of course, the analogue of "tree admittance product" is "impedance product of all but one elements in a cutset."

The essential capacitors will be labeled C_1 through C_{n1}, the essential inductors will be labeled $L_{(n1 + 1)}$ through L_n. The excess capacitors will be labeled $C_{(n + 1)}$ through C_{n2} and the excess inductors will be labeled $L_{(n2 + 1)}$ through L_{n3}. Observe that n is the number of state variables. The state variables associated with capacitor C_m or inductor L_k will be denoted by x_m or x_k, respectively, m and k being less than n.

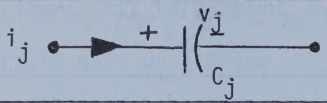
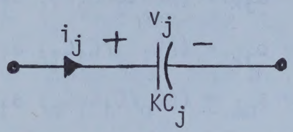
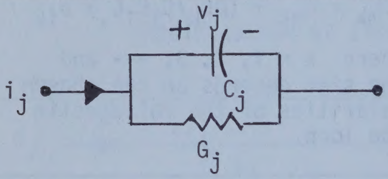
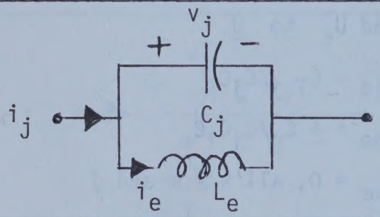
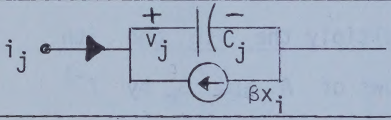
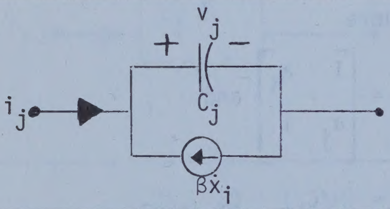
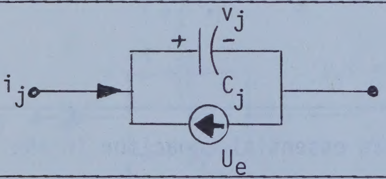
TABLE I
Modification of Essential Reactive Elements

		Changes in A and B Matrices
1.		$a'_{jk} = (C_T/C'_T)\,a_{jk}$ $b'_{jk} = (C_T/C'_T)\,b_{jk}$ for $k = 1, 2, 3, ---$ Refer to "Derivation for Compensation" for $a'_{mk}{}^*$ and b'^{*}_{mk}.
2.		$a'_{jj} = a_{jj} - (C_{T,j}/C_jV_T)G_j$ $a'_{mj} = a_{mj} \pm (C_p/C_mV_jC_T)G_j$ where the sign is the same as that of a_{mj}
3.		Add i_e to $\underline{x}$, a column e to the A matrix and a row e, of zeros, to the B matrix. $a'_{ke} = 0$, all $k \neq m$ and j $a'_{je} = -C_{T},j/C_jC_T$ $a'_{me} = \pm C_p/C_jC_mC_T$ $a'_{ej} = 1/L_e$ $a'_{jj} = a_{jj}$

4.	 v_j + − i_j C_j βx_i	$a'_{ji} = a_{ji} + \beta C_{T,j}/C_jC_T$ $a'_{mi} = a_{mi} \pm \beta C_p/C_jC_mC_T$
5.	 v_j + − i_j C_j βx_i	$a'_{jk} = a_{jk} + (\beta C_{T,j}/C_jC_T)\, a_{ik}$ $b'_{jk} = b_{jk} + (\beta C_{T,j}/C_jC_T)\, a_{ik}$ $a'_{mk} = a_{mk} \pm (\beta C_p/C_jC_TC_m)\, a_{ik}$ $b'_{mk} = b_{mk} + (\beta C_p/C_jC_TC_,)\, b_{ik}$ where $k = 1, 2, 3$, --- and the sign depends on the chosen polarities of the voltages in the loop.
6.		Add U_e to $\underline{U}$. $b_{je} = C_{T,j}/C_jC_T$ $b_{me} = \pm C_p/C_jC_TC_m$ $b_{ke} = 0$, all $k \neq m$ and j
7.	 i_j v_j + − C_j $M\dot{v}_i$ $M\dot{v}_j$ C_i i_i + v_i −	Multiply the ith and jth rows of A and B by T^{-1} where $T = \begin{bmatrix} 1 & a_i \\ a_j & 1 \end{bmatrix}$ and $a_k = (M/C_k) \cdot C_{T,k}/C_T$ $k = i,j$

*m, anywhere in this Table, corresponds to an essential capacitor in the loop C_{Lj}.

TABLE II
Modification of Excess Reactive Elements

		Changes in A and B Matrices
1.		a'_{mk}* and b'_{mk}* will be evaluated, individually, using superposition, k = 1, 2, 3, ---
2.		$a'_{mk} = a_{mk} \pm (C_p/C_mC_jC_T)\cdot G_j$ where k is such that C_k belongs to C_{Lj}.
3.	 v_j + − i_j C_j i_e L_e	Add i_e to $\underline{x}$, a column e to the A matrix and a row e, of zeros, to the B matrix. $a'_{ke} = 0$, all $k \neq m$ $a'_{me} = \pm C_p/C_jC_mC_T$
4.	 + − i_j v_j C_j βx_i	$a'_{mi} = a_{mi} \pm \beta C_p/C_jC_mC_T$
5.	 v_j + − i_j C_j $\beta \dot{x}_i$	$a'_{mk} = a_{mk} \pm \beta(C_p/C_jC_mC_T)a_{ik}$ $b'_{mk} = b_{mk} \pm \beta(C_p/C_jC_mC_T)b_{ik}$ where k = 1, 2, 3, --- and the sign depends on the chosen polarities of the voltages in the loop.
6.		Add U_e to $\underline{U}$. $b_{me} = \pm C_p/C_jC_mC_T$ $b_{ke} = 0$, all $k \neq m$

* m, anywhere in this Table, corresponds to an essential capacitor in the loop C_{Lj}.

DERIVATION OF COMPENSATIONS

The compensations of the entries of A and B matrices due to augmentation of essential reactive elements are described briefly in Table I. Derivation of these modifications will be given next. Consider case (1) in the Table. Let a_{jk} be an entry of the A matrix before augmentation.

$$a_{jk} = \dot{v}_j/x_k \;\Big|\; \text{all other state variables and independent sources are deactivated}$$

Observe that the currents entering the capacitor loop C_{Lj} are linearly related to the excited variable x_k, and the current i_j itself is linearly related to the currents entering the capacitor loop externally [2]. In view of these facts and the lemma in the appendix, we have

$$\begin{aligned} i_j &= P_{jk} \cdot (C_j/C_T)\, x_k \\ a_{jk} &= \dot{v}_j/x_k \\ &= 1/C_j \cdot (P_{jk}) \cdot C_j/C_T \\ &= P_{jk} \cdot 1/C_T,\ k = 1, 2, 3, --- \end{aligned}$$

where P_{jk} is a constant independent of C_j.

$$a'_{jk} = P_{jk} \cdot 1/C'_T$$

$$a'_{jk} = (C_T/C'_T)\, a_{jk},\ k = 1, 2, --- \qquad (1a)$$

where the unprimed and primed parameters belong to the original and augmented networks, respectively.

Similarly one can show that

$$b'_{jk} = (C_T/C'_T)\, b_{jk},\ k = 1, 2, --- \qquad (1b)$$

For $m \neq j$ such that C_m belongs to C_{Lj}, a'_{mk} and b'_{mk} depend on the topology of the network as well as a_{mk} and b_{mk}, respectively. They can be obtained, individually, using superposition [2].

(2) When a conductance G_j is connected across C_j, its effect is to cause a current v_jG_j to enter the loop C_{Lj} at one of the nodes and leave it at some other node when v_j is activated and other state variables and independent sources are deactivated. The fraction of this current that passes through C_j is given by

$$(C_{T,j}/C_T) \cdot v_jG_j \qquad \text{and}$$

$$i'_j = i_j - (C_{T,j}/C_T)\, G_j v_j$$

or
$$C_j \dot{v}'_j = C_j \dot{v}_j - (C_{T,j}/C_T) \cdot G_j v_j$$

$$a'_{jj} = a_{jj} - (C_{T,}j/C_j C_T)\ .\ G_j \tag{2a}$$

For $m \neq j$ such that C_m belongs to the loop C_{Lj},

$$a'_{mj} = a_{mj} \pm (C_p/C_m C_j C_T)\ .\ G_j \tag{2b}$$

where the sign is the same as that of a_{mj}.

(3) Addition of an inductor as shown in the Table adds a new column to the A matrix, element i_e to the state vector $\underline{x}$ and a row of zero elements to the B matrix.

$$a'_{ke} = 0, \text{ all } k \neq m \text{ and } j \tag{3a}$$

$$a'_{je} = -C_{T},_{j}/C_j C_T \tag{3b}$$

For $m \neq j$, such that C_m belongs to the loop $C_L j$,

$$a'_{me} = \pm\, C_p/C_j C_m C_T \tag{3c}$$

Also,

$$a'_{ej} = 1/L_e \tag{3d}$$

$$a'_{jj} = a_{jj} \tag{3e}$$

(4) When the state variable x_i is excited, a current source x_i pumps current into the loop $C_L j$.

$$a'_{ji} = a_{ji} + \beta C_{T},_{j}/C_j C_T \tag{4a}$$

For $m \neq j$ such that C_m belongs to the loop C_{Lj},

$$a'_{mi} = a_{mi} \pm \beta C_p/C_j C_m C_T \tag{4b}$$

(5) Assume, in this case, that the state variable x_i is not associated with a capacitor in a capacitor loop or an inductor in an inductor cutset.

Assume, in case 7, that C_i and C_j belong to non-intersecting capacitor loops. The changes in this case can be obtained by exciting v_i and v_j simultaneously (while deactivating other state variables and sources) and evaluating i_i and i_j in terms of v_i and v_j.

Table II gives changes in the entries of A and B matrices, when the reactive elements being modified are excess elements. These can be derived in a way similar to the essential reactive element cases discussed previously.

CONCLUSION

Compensations in the entries of the system matrix A and the source distribution matrix B due to various types of modifications in improper networks have been derived and tabulated in this short paper. These compensations are very similar to those in the corresponding cases for proper networks, and can be computed with ease by inspection in a time much shorter than that required for obtaining the state equations for the augmented network without taking advantage of the already found equations of the original network. The results obtained can be useful in analysis, synthesis and study related to sensitivity and effect of parasitic elements.

APPENDIX

Lemma: Let $C_{m1}, C_{m2}, ---, C_{mk}, C_{n1}, C_{n2}, ---, C_{n\ell}$ be $(k + \ell)$ capacitors in a capacitor loop. Let I be the current entering the loop externally and leaving as shown in Fig. 1. The current i_{mj}, $1<j<k$, in the capacitor C_{mj} can be expressed as

$$i_{mj} = \frac{C_{T,\, m_1,\, m_2,\, ---,\, m_k}}{C_T} . I$$

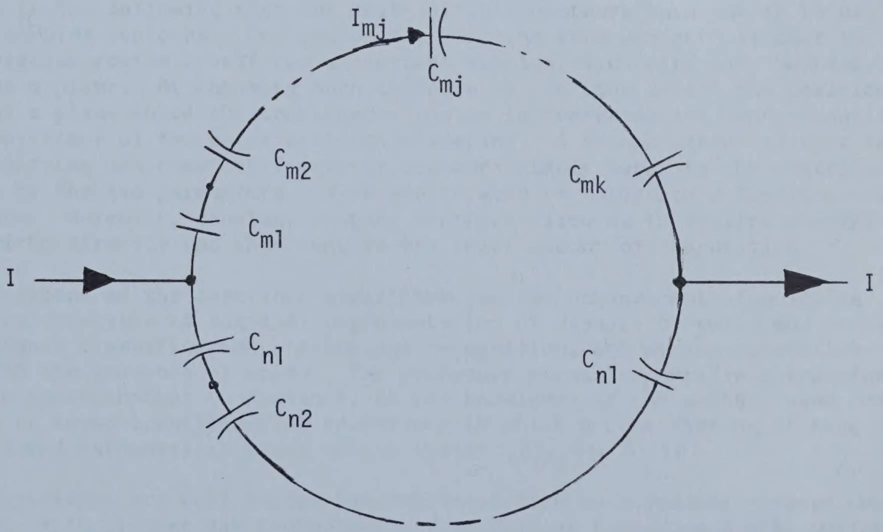


Figure 1

REFERENCES

(1) M. F. Moad, "Compensation of Network State Equations Due to Modification of Reactive Elements", IEEE Transactions on Circuit Theory, Vol. CT-19, Number 1, January, 1972 pp. 66-68.

(2) Moiez A. Tapia, "State Equations for an Improper Network Without Having to Invert Matrices", IEEE Transactions on Circuit Theory Vol. CT-20, No. 4, July 1973, pp. 420-422.

(3) B. C. Kuo, Linear Networks and Systems. New York: McGraw-Hill, 1967.

(4) R. A. Rohrer, Circuit Theory: An Introduction to the State Variable Approach. New York: McGraw-Hill, 1970.

AN ALGORITHM FOR COMPUTING LAPLACE AND Z TRANSFORM OF A TIME SERIES WITH APPLICATION TO SYSTEM IDENTIFICATION

MICHAEL J. CORINTHIOS
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL, MONTRÉAL, CANADA.

ABSTRACT

Algorithms for computing poles and zeros of a system's transfer function through Z transformation, of a finite time series are introduced. The discrete Laplace or z transform algorithms to be described are shown to utilize the fast Fourier transform algorithm and thus yield an appreciable saving in computation time. It is shown that a judicious choice of the contour of transformation in the complex plane and of the frequency resolution reduces considerably the amount of computation, even when compared with such fast algorithms as the Chirp z transform (CZT) [1],[2]. A procedure for on-line identification of a system through a direct computation of the poles of its transfer function are described.

I. INTRODUCTION

Given a linear system, the poles and zeros of its transfer function appear as positive or negative peaks in the complex Laplace or z plane. A logical way that may be proposed for locating the singularities and meanwhile obtain a general view of the system's behavior is to compute the discrete Laplace or z transform of the time series that is a finite duration of its impulse response and identify the points in the plane where such high peaks occur.

Having decided to follow this approach we need to choose suitable contours along which to compute the z transform of the given time series. It is shown in the following that the most suitable contours turn out to be constant damping contours. One reason for choosing such contours is that the convergence region itself has a constant damping, i.e. circular, boundary in the z plane. By choosing such contours we can thus detect the position in the z plane where the convergence region is traversed and hence identify the positions of the poles with least damping. A second reason is that constant damping and constant frequency contours simply separate the contributions of the two parameters, as we desire when we integrate a function over a plane. Moreover, constant damping contours allow us to utilize the FFT algorithm directly and thus require the least amount of computation.

Applications of the described algorithms include enhancement of poles in spectral analysis of signals, representation of signals by poles and zeros for signal classification, coding and recognition, and system identification in the presence of noise. The procedure presented here is a transform domain approach that is distinct, to the knowledge of the author, when compared to known identification techniques in which a best fitting is made to an assumed mathematical model of the system [3],[4],[5],[6].

The algorithms are well suited for implementation on a special purpose computer. With present day technology a fast Fourier transform can be performed in real time on signals sampled at frequencies in the megahertz range [7]-[10]. In [10] the author describes a real time FFT processor which was constructed and performs a 256-point transform in 128 μsec. A 64-point transform can be computed in real time for signals sampled at close to 10 MHz. A system of the tenth order can thus be identified on line at sampling frequencies between 500 KHz and 1MHz. Identification of relatively rapidly time-varient systems is thus possible.

II. PROBLEM DEFINITION

The z transform of a time series f_n is given by

$$F(z) = \sum_{n=o}^{\infty} f_n z^{-n} \quad (1)$$

If $z = e^{sT} = e^{(\sigma + j\omega)T}$ then $F(z) = \sum_{n=o}^{\infty} f_n e^{-nsT}$

which is the discrete Laplace transform of the time series.

$$F(z) = \sum_{n=o}^{\infty} f_n e^{-n\sigma T} e^{-jn\omega T}$$

For a finite time series of N points we have

$$F(z) = \sum_{n=o}^{N-1} f_n e^{-n\sigma T} e^{-jn\omega T} \quad (2)$$

where T is the sampling interval in the time domain. We also have

$\omega = k\Delta\omega$ where $\Delta\omega$ in the sampling frequency interval

& $\sigma = i\Delta\sigma$ where $\Delta\sigma$ is the sampling interval along the σ axis

If the transform is to be evaluation at M points so that

$$\Delta\omega = \frac{2\pi}{MT}$$

We can thus rewrite Eq. (2) in the form

$$F(e^{(i\Delta\sigma + jk\Delta\omega)T}) = \sum_{n=o}^{N-1} f_n e^{-j\frac{nk}{M} x2\pi} e^{-ni\Delta\sigma T} \quad (3)$$

$$\text{i.e. } F(i\Delta\sigma, jk\Delta\omega) \triangleq F(e^{(i\Delta\sigma + jk\Delta\omega)T}) = \sum_{n=o}^{N-1} f_n a^{-ni} w^{-nk} \quad (4)$$

where

$$a = e^{\Delta\sigma T} \quad (5)$$

$$w = e^{-j\frac{2\pi}{M}} \quad (6)$$

If we assume that we are interested in M transform points in the z domain we have a two dimensional function to evaluate with a number of operations NxM where an operation is a complex multiplication by $a^{-ni} w^{-nk}$ followed by addition. Scanning the z plane uniformly by $M=N^2$ point would call for N^3 such operations, which is computationally excessive.

To reduce the number of computations Rabiner et al introduced the Chirp Z transform (CZT) which makes use of the Fast Fourier Transform FFT algorithm to evaluate Eq. (1) over a contour in the z plane.

The contour is chosen to be a spiral or, as a special case, a circle converging to or diverging from the plane's centre and starting at any point in the z plane.

Assuming that the spiral starts in the z plane at the point z_o given by

$$z_o = e^{s_oT} = e^{\sigma_oT} e^{j\omega_oT}.$$

The Kth point z_k is given by

$$z_k = e^{\sigma_oT} e^{k\Delta\sigma T} e^{j\omega_oT} e^{jk\Delta\omega T} \quad (7)$$

Letting

$$A_o = a = e^{\sigma_o T}, \; W_o = e^{-\Delta\sigma T} \quad (8)$$

then we can write z_k in the form

$$z_k = A_o e^{j\omega_o T} \quad w_o^{-k} \, e^{jk\Delta\omega T} \tag{9}$$

$$= AW^{-k} \tag{10}$$

where

$$A = A_o e^{j\omega_o T} \qquad = e^{(\sigma_o + j\omega_o)T} \tag{11}$$

and

$$W = W_o e^{-j\Delta\omega T} \qquad = W_o \; e^{-j2\pi/M} \tag{12}$$

$$= W_o w$$

The transform along such a contour would thus take the form

$$F(\sigma_o + k\Delta\sigma, \omega_o + k\Delta\omega) = \sum_{n=o}^{N-1} f_n \, A^{-n} \, W^{nk} \quad ; \quad k = 0,1,\ldots,M-1 \tag{13}$$

Rabiner et al utilized the relation

$$nk = [n^2 + k^2 - (n-k)^2]/2$$

to change the product of powers into sums. The CZT then computes M transform points from a time series of N points, where M and N are arbitrary. It calls for weighting the input samples by $A^{-n} W^{n^2/2}$, performing an (N+M) point convolution and weighting the resulting M points by $W^{k^2/2}$; $k=0,1,\ldots M-1$.

As Rabiner et al report there arise situations in which the first weighting operation may cause overflow or underflow, and the accumulated computational errors due to these weighting operations may distort the computed spectrum. A more important reason for such problems with overflow could be attributed to the fact that the series of summation tends toward a divergent one once the boundary of convergence has been traversed by the spiral contour. In the following we show that the need for convolution is eliminated if the transformation is performed over a constant damping contour with a properly chosen frequency resolution that is related to the radix of factorization of the FFT algorithm. Transformation over constant-frequency contours is performed through a direct evaluation of (2) and helps us obtain higher precision in computing the damping coefficient. The computation involved in transformation over each of these contours is first described. Subsequently, a procedure for locating the singularities is given.

III. CONSTANT-FREQUENCY AND CONSTANT DAMPING CONTOURS

3.1 Constant-Frequency Contours

For a constant-frequency (radial) contour we have

$$\Delta\omega = 0 \tag{14}$$

The k th point z_k is given by

$$z_k = e^{(\sigma_o + j\omega)T} \; e^{k\Delta\sigma T} \tag{15}$$

$$= AW_o^{-k} \tag{16}$$

and we have

$$F(z_k) = F(\sigma_o+k\Delta\sigma,\omega_o) = \sum_{n=0}^{N-1} f_n A^{-n} W_o^{nk}; \tag{17}$$

$$k = 0,1,\ldots,m_\sigma-1 \tag{18}$$

If we, therefore modify our time series by weighting it by A^{-n} to obtain the modified time series g_n given by:

$$g_n = f_n A^{-n} \tag{19}$$

Then we can write

$$F(z_k) = \sum_{n=0}^{N-1} g_n W_o^{nk} \tag{20}$$

$$= \sum_{n=0}^{N-1} g_n e^{-nk\Delta\sigma T} \quad ; \quad k = 0,1,\ldots, M_\sigma-1 \tag{21}$$

Note that the weighting coefficients in Eq(21) are real valued. We can write Eq(21) in the form

$$f(z_k) = \sum_{n=0}^{N-1} g_n W_o^{nk}$$

The real array W_o^{nk} can be initially computed and stored as the array

$$V(m) = W_o^m = e^{-m\Delta\sigma T}$$

and utilized for summation over any radial contour; with $\Delta\sigma$ kept the same for the different radial lines. As will be shown subsequently the value of M_σ need not be large. The values M_σ , σ_o and $\Delta\sigma$ need not bear any relation to the number of samples N of the time series.

3.2 Constant-Damping Contours

For a constant-damping circular contour we have

$$\Delta\sigma = 0 \tag{24}$$

and therefore

$$z_k = e^{(\sigma_o+j\omega_o)T} \qquad e^{jk\Delta\omega T} \tag{25}$$

The transform along this contour is thus

$$F(\sigma_o,k\Delta\omega) = \sum_{n=0}^{N-1} f_n \quad e^{-n(\sigma_o+j\omega_o)T} \quad e^{-j2\pi nk/M_\omega} \tag{26}$$

$$F(\sigma_o,k\Delta\omega) = \sum_{n=0}^{N-1} f_n \; A^{-n} \; w^{nk} \tag{27}$$

$$= \sum_{n=0}^{N-1} g_n \; w^{nk} \tag{28}$$

where

$$g_n = f_n A^{-n} = f_n \quad e^{-n(\sigma_o+j\omega_o)T} \tag{29}$$

$F(\sigma_0, k\Delta\omega)$ is thus simply the discrete Fourier transform of the modified series g_n. If the number of transform points M_ω is chosen equal to N then $F(\sigma_0, k\Delta\omega)$ can be computed by Fourier transforming the time series g_n using the classical FFT algorithm. The number of operations thus required is equal to $N\log_r N$, where r is the radix of factorization of the discrete Fourier transform.

If we choose a number of transform points M_ω different from the number of points N of the time series then a slight modification of the "classical" form of the FFT algorithm is needed for the efficient computation of the discrete Fourier transform. These modified algorithms are particularly useful when the frequency resolution is several times higher than the "normal" resolution in which N uniformly spaced transform points cover the whole frequency spectrum [7]. It may be noted, that the only restriction on the number of transform points M_ω is that M_ω/N should be a power (positive or negative) of the radix r of factorization of the discrete Fourier transform. It is also noted that if the number of points M_ω is inversely proportional to the resolution of the transformation, which is usually the case, then the number of operations required has as upper bound the number of operations called for if the classical FFT algorithm is used.

IV. PROCEDURE FOR LOCATING POLES AND ZEROS

Having considered the computations involved in transformation along radial and circular contours let us consider the following procedure for locating the poles and zeros of the given time series.

We may start by a transformation involving a contour on the unit circle. The amplitude spectrum thus obtained can be used to locate the frequencies of the nearest poles of the system. Radial-contours along and close to these peak frequencies are thus chosen for transformation. By comparing phase values at the neighbourhood of the peak-frequency radial contours a sign inversion locates the damping of the first poles; those of least damping. This process can be repeated until the first poles are located with the required precision. The first poles are then deleted and the order of the system reduced. This process is repeated until all significant poles have been computed.

V. PRACTICAL CONSIDERATIONS

There are numerous considerations which affect the choice of the number of transform points M_σ and M_ω. In the following we summarize these considerations.

5.1 Contours on Complete Circles

A special case of a circular contour is one in which the contour is a complete (or half) circle and where the number of transform points M_ω is equal to N. This is the case of "normal" frequency resolution and is the simplest case. It is always useful in scanning circularly the z domain. Moreover, the number of computations involved here constitutes an upper bound on the number of operations required for transformation on a part of a circle with the same or less resolution. Let us therefore rewrite the eq. (26) for circular contours for this particular case.

$$M_{\omega} = N \tag{30}$$

and

$$\omega_{o} = 0 \tag{31}$$

We have

$$F(\sigma_{o},k\Delta\omega) = \sum_{n=o}^{N-1} f_{n} e^{-n\sigma_{o}T} e^{-j2\pi\, nk/N} \tag{32}$$

$$= \sum_{n=o}^{N-1} g_{n} e^{-j2\pi nk/N} \tag{33}$$

where

$$g_{n} = f_{n} e^{-n\sigma_{o}T} \; ; \; n = 0,1,\ldots,N-1 \tag{34}$$

We notice therefore that the modified series g_n is obtained through weighting the original time series f_n by the real values $e^{-n\sigma_o T}$. If an operation designates multiplication of two complex words then the number of operations involved in weighting the time series f_n to obtain g_n is slightly less than N/4.

Having obtained the time series g_n we next apply the FFT algorithm in its normal form to obtain the transform $F(\sigma_o\ , k\Delta\omega)$. This requires $N\log_r N$ rations. The total number of operations on a completely circular contour denoted N_{cir}, is therefore given by

$$N_{cir} = N/4 + N\log_{r}N$$

operations.

From the above it may be shown that the constant-damping transformation described here requires much less computation than the CZT algorithm. Moreover, the described algorithm does not call for weighting by $W^{-n^2/2}$ and thus no problems with overflow, underflow or excessive accumulated errors are encountered in computing the transform. It should be emphasized that the efficiency of the CZT is unquestionable when transforms are computed along spiral contours. Searching for poles and zeros is simplified however if the z plane is scanned by circles radiating and radial lines emanating from the center of the plane.

5.2 Contours on Circle Segments

With the same normal resolution, if the contour is part of a circle which starts at zero frequency the computations involved are less than found in Sec. 5.1 above. If the circle segment does not start at zero frequency ω_o then the time series f_n is weighted by a complex exponential, according to Eq. (29), then transformed using the Fast Fourier transform algorithm according to Eq. (28).

If a higher resolution is needed then the number of operations involved in obtaining the modified times series is the same as above and the number of operations required for a high-resolution FFT is proportional to $RN\log_r N$, where the number of transform point of interest is N points and the frequency resolution is R times the normal resolution. By R times the normal resolution we mean that the sampling interval is ω_o/R.

If we choose to start our algorithm by a coarse resolution circular transform then we need to perform first an aliasing of the modified time series before we apply the FFT algorithm. A smaller FFT is thus needed. Thus if we need to perform a circular transform with R times coarser resolution the number of operations of the FFT becomes approximately $(N/R)\log_r(N/R)$ operations.

5.3 Radial line Contours

As has been shown above (Eqs.(20) to (24)) summation over radial contours calls for weighting the original time series to obtain the modified time series g_n. This time series is then weighted by a real exponential $W_o^{nk} = e^{-nk\Delta\sigma T}$. The number of operations involved here is N/2 per transform point. For M_σ points we require $NM_\sigma/2$ operations. However, the number of points M_σ along radial lines is usually small and located close to a pole or zero. Thus the number of operations will usually be less than that required for summation over circular contours.

VI. REAL TIME PROCESSING

Fig. 1 shows a block diagram of a system for on-line system identification by radial-circular Z transformation. The preweighting unit applies the weighting coefficients A^{-n} to the input time series f_n and produces the modified time series. The Contour Computation unit applies either a radial or a circular transformation at any point in time. The output of this unit is used by the Contour Decision unit to set the subsequent contour and its resolution.

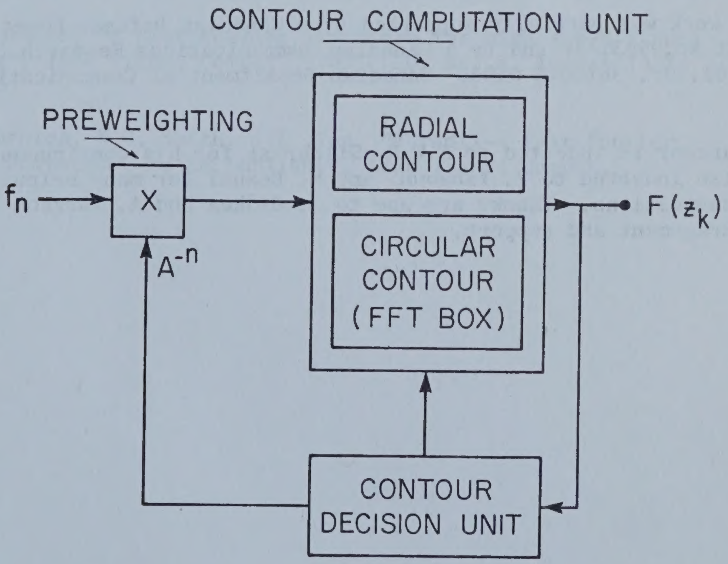


Fig. 1: ON-LINE IDENTIFICATION SYSTEM

The amount of computation involved in locating all the poles and zeros of a system is linearly proportional to the number of poles and zeros of the system. A total number close to K times the number of operations involved in computing one FFT on N points will be needed to identify a system with K poles and zeros. With present day technology a fast Fourier transformation can be performed in real time on signals sampled at frequencies in the Megahertz range [8 to 10]. In [10] the author describes a real time FFT processor which has been constructed and performs a 256-point transform in 128μsec. A 64-point transform can be computed in real time for signals sampled at close to 10MHZ. A system with 10 poles and zeros can thus be identified on line at sampling frequencies close to 500KHZ. Identification of relatively rapidly time-variant systems is thus presently possible. These same considerations apply to real-time signal classification in terms of the locations of its poles and zeros.

VII. EXAMPLE

As an illustration the amplitude and phase of constant-damping transforms for a system of the twentieth order is shown in Fig. 2. The first curve shows the transform on the unit circle. The second displays the transform on a contour lying between the unit circle and the first poles. The transform on a contour passing through the third pole is shown as the third curve. The fourth curve shows the result of transformation beyond the contour of convergence. The amplitude curves display the increasing peakedness of the amplitude spectrum when the poles are approached from the unit circle. The phase curves shows a reversal in slope when the pole is traversed

ACKNOWLEDGEMENT

This work was partially supported by a Canadian Defense Research Board Grant No. 9931-36 and by a Canadian Communications Research Centre Contract No. 01. Gr. 36100-2-0195, Canadian Department of Communications Ottawa.

The author is indebted to A.W.R. Gilchrist for his continuous support. He is also indebted to P. Yansouni and M. Beshai for many helpul discussions and suggestions. Thanks are due to J. O'Shea and A. Turgeon for their encouragement and support.

REFERENCES

[1] Rabiner, L.R., Schafer, R.W. and Rader, C.M., "The Chirp Z-Transform and Its Application", Bell System Technical Journal, May-June, 1969, pp. 1249-1291.

[2] Rabiner, L.R., Schafer, R.W. and Rader, C.M., "The Chirp Z-Transform Algorithm" IEEE Trans. Aud. and Electroacoust, Vol Au-17, No. 2, 1969 pp. 86-96.

[3] R.E. Kalman, "Design of a self-optimizing control system", Trans. ASME, Vol. 80, pp. 468-478, February, 1958, pp. 461-464.

[4] Y.C. Ho and B.H. Whalen "An Approach to the identification and Control on Linear Dynamic Systems with Unknown Parameters, IEEE Trans. Automatic Control, July 1963, pp. 255-256.

[5] K.E. Bollinger and J.C. Mathur, "To compute the zeros of large systems", IEEE Trans. Automatic Control, Feb 1971, pp 95-96.

[6] A.B. Turgeon, "Identification en temps réel des systèmes échantillonnés en présence de bruits de mesure", Automatisme, Vol. 15, No. 3, pp 98-103 mars 1970.

[7] M.J. Corinthios and Y. Geadah, "Fourier transformation with variable bandwidth and frequency resolution", Rapport technique EP73-R-3, Ecole Polytechnique de l'Université de Montréal, Jan 1973, pp 1-15, to be published.

[8] M.J. Corinthios, "The design of a class of fast Fourier transform computers", IEEE Trans. Comput, Vol. C-20, June 1971, pp 617-623.

[9] M.J. Corinthios, "A Fast Fourier Transform for high speed signal processing", IEEE Trans. Comput., Vol C-20, No. 8, August 1971, pp 843-846.

[10] M.J. Corinthios, K.C. Smith, J.L. Yen, "A radix-4 fast Fourier transform Computer", to be published, IEEE Trans. Comput.

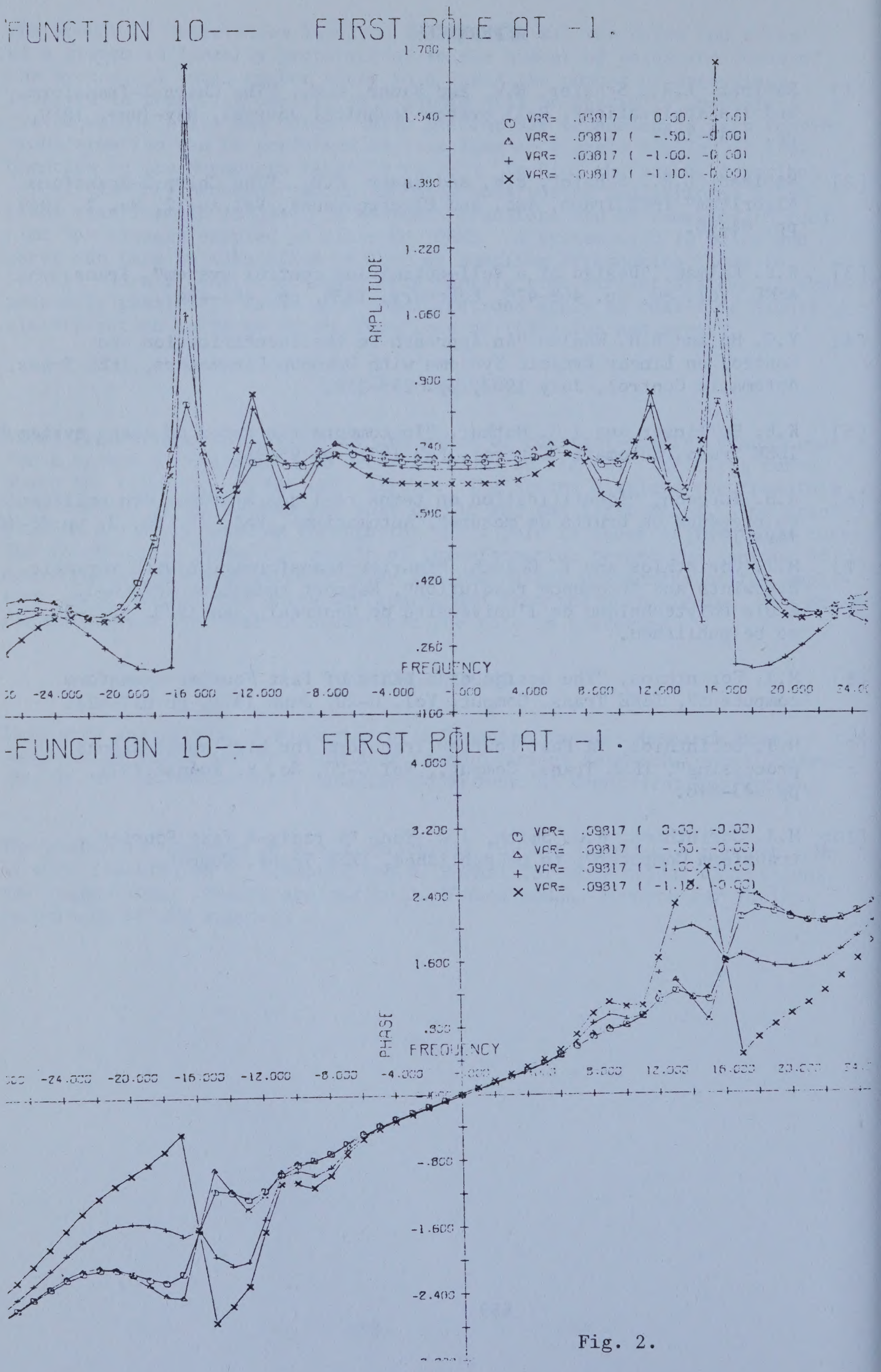

FUNCTION 10--- FIRST POLE AT -1.
AMPLITUDE
1.700
1.540
1.380
1.220
1.060
.900
.740
.580
.420
.260
.100
VAR= .09817 (0.00, -0.00)
VAR= .09817 (-.50, -0.00)
VAR= .09817 (-1.00, -0.00)
VAR= .09817 (-1.10, -0.00)
FREQUENCY
-24.000
-20.000
-16.000
-12.000
-8.000
-4.000
.000
4.000
8.000
12.000
16.000
20.000
FUNCTION 10--- FIRST POLE AT -1.
PHASE
4.000
3.200
2.400
1.600
.800
-.800
-1.600
-2.400
VAR= .09817 (0.00, -0.00)
VAR= .09817 (-.50, -0.00)
VAR= .09817 (-1.00, -0.00)
VAR= .09817 (-1.10, -0.00)
FREQUENCY
-24.000
-20.000
-16.000
-12.000
-8.000
-4.000
.000
8.000
12.000
16.000
20.000

Fig. 2.

NOVEL APPLICATION OF THE NUMERICAL LAPLACE TRANSFORM INVERSION

M. NAKHLA(MRS.), K. SINGHAL AND J. VLACH
University of Waterloo,
Waterloo, Ontario, Canada

ABSTRACT

A method published in [1,2,3] and programs published in [2,4] can be used for several novel applications. The examples and discussion consider time-responses of linear networks.

Consider a function F(s) to be inverted into f(t) numerically. The inversion formula is

$$\hat{f}(t) = \frac{-1}{t} \sum_{i=1}^{M} K_i \; F\left(\frac{z_i}{t}\right) \tag{1}$$

where $\hat{f}(t)$ is an approximation to f(t) and K_i, z_i are complex numbers

Formula (1) was obtained from

$$f(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(s)\, e^{st}\, ds \tag{2}$$

by first instroducing a new variable z=st and then replacing e^z by its Padé approximation with N being the degree of the numerator and M that of the denominator. In other words

$$e^z \approx \xi_{N,M}(z) = \frac{\sum\limits_{i=o}^{N} a_i z^i}{\prod\limits_{i=1}^{M} (z-z_i)} = \sum_{i=1}^{M} \frac{K_i}{z-z_i} \tag{3}$$

Lemma 1

The Padé approximation $\xi_{N,M}(z)$ must be chosen such that the function F(z/t). $\xi_{N,M}(z)$ has at least two more poles than zeros.

The proof follows from residue calculus. The infinite arc, closing the path of integration, does not contribute to the value of the integral if the integrated function has two more poles than zeros.

Remark: In network applications, a realizable transfer function F(s) can have at most as many poles as (finite) zeros. It is, therefore, a safe practice to use $\xi_{N,M}(z)$ with N=M-2. The tables are provided for N=M,M-1,M-2 [4]. For other situations, the program described in [2] can be used.

Lemma 2

Formula (1) inverts exactly the functions

$$F_m(s) = s^{-m} \qquad N+2-M \le m \le M+N+1 \tag{4}$$

The inverse is equal to

$$f_m(t) = \hat{f}_m(t) = \frac{t^{m-1}}{(m-1)!} \qquad (5)$$

Note that negative m can also appear in (4), (5), i.e. the method is valid for the delta function and its derivatives. In such cases f(t)=0 for positive t since, by definition [5], 1/n! = 0 for negative n.

The proof has two parts. For $m > 0$ it is known that $\mathcal{L}^{-1}(s^{-m}) = \frac{t^{m-1}}{(m-1)!}$. Inserting s^{-m} for F in (2), integrating along Re $z = c$ and closing the contour in the left half plane, we get

$$\hat{f}_m(t) = \frac{t^{m-1}}{(m-1)!}\left[\frac{d^{m-1}}{dz^{m-1}} \xi_{N,M}(z)\right]_{z=0}$$

where the formula for calculating the residue of the m-th order pole at the origin was used. However, $\xi_{N,M}(z)$ is a Padé approximation to e^z having the first M+N+1 terms of its Taylor expansion identical to the Taylor expansion of e^z. Therefore

$$\frac{d^m}{dz^{m-1}} (\xi_{N,M}(z))_{z=0} = 1 \qquad m=1,2,\ldots M+N+1$$

and

$$f_m(t) = \hat{f}_m(t) = \frac{t^{m-1}}{(m-1)!} \qquad 0 < m \le M+N+1$$

The second part of the proof is for $N+2-M \le m \le 0$. Consider such a choice of N,M that m is nonpositive and let $n = -m$. Then the test function is $F(s) = s^n$ and (2) has the form

$$\hat{f}_m(t) = \frac{1}{2\pi jt} \int_{c'-j\infty}^{c'+j\infty} \left(\frac{z}{t}\right)^n \xi_{N,M}(z)dz = \frac{t^{-n-1}}{2\pi j} \int_{c'-\infty}^{c'+\infty} z^n \xi_{N,M}(z)dz$$

Because of the choice of m, the function under the integral sign has no poles in the left half plane, the numerator is lower than the denominator by at least two degrees, and thus the integral is equal to zero when the path is closed in the *left* half plane. Combining with (1)

$$\hat{f}_m(t) = -t^{-n-1} \sum_{i=1}^{M} K_i z_i^n = 0 \qquad \text{for } t > 0, \; N+2-M \le m \le 0.$$

The first part of the proof was given by Piessens [6]. The second part is new.

The proof has direct use in checking the correctness of the values of the residues and poles. Insert t =1 and $F(z/t) = (z/t)^{-m}$ into (1). Then

$$-\sum_{i=1}^{M} \frac{K_i}{z_i^m} = \frac{1}{(m-1)!} \qquad \text{for } N+2-M \le m \le M+N+1 \qquad (6)$$

If, for instance, M=10 and N=6, equation (6) shows that

$$\sum_{i=1}^{M} K_i z_i^2 = \sum_{i=1}^{M} K_i z_i = \sum_{i=1}^{M} K_i = 0$$

$$-\sum_{i=1}^{M} \frac{K_i}{z_m^i} = \frac{1}{(m-1)!} \qquad \text{for } m=1,2,.....17$$

Lemma 3

Formula (1) exactly inverts the first M+N+1 terms of the Taylor expansion of any time response.

Proof: Follows directly from Lemma 2. Consider a time response

$$f(t) = \sum_{i=o}^{\infty} \frac{a_i}{i!} t^i = \sum_{i=o}^{M+N} \frac{a_i}{i!} t^i + R(t) = f_T(t) + R(t)$$

where R(t) are all remaining terms of the expansion. The Laplace transform of the truncated part is equal to

$$F_T(s) = \sum_{i=o}^{M+N} \frac{a_i}{s^{i+1}}$$

Since each term of this sum is inverted exactly, the whole sum is also, which concludes the proof.

Using (6), we next show that formula (1) is correct for $t \to 0$ if f(0) is bounded and $M \geq N+1$. For the sake of simplicity, we will explain this on an example. Consider $F(s) = s/(s^2+a)$ which, by the initial value theorem, is $\lim_{t\to 0} f(t) = \lim_{s\to\infty} sF(s) = 1$. Inserting into (1), we have

$$-\lim_{t\to 0} \frac{1}{t} \sum_{i=1}^{M} K_i \frac{\frac{K_i}{t}}{\left(\frac{z_i}{t}\right)^2 + a} = -\sum_{i=1}^{M} \frac{K_i}{z_i} = 1$$

where the last step follows from (6). Graphs of errors given in [4] confirm this phenomena for other functions.

Derivatives of the time response can be obtained using

$$[f^{(n)}(t)] = s^n F(s) - \sum_{i=1}^{n} f^{(i-1)}(0+)\, s^{n-i} .$$

Choose N,M such that $s^n F(s)\, \xi_{N,M}(z)$ has at least two more zeros than poles. Then all terms after the $\sum$ operator in (7) will be zero, due to (6) and the n-th derivative is inverted as

$$\hat{f}^{(n)}(t) = \frac{-1}{t} \sum_{i=1}^{M} K_i \left(\frac{z_i}{t}\right)^n F\left(\frac{z_i}{t}\right) \qquad (7)$$

Note that integrals of the time function do not present any problem, since the number of poles increases and thus improves the situations.

Applications

Responses of networks are sometimes defined by such technical quantities as "rise time", "time delay","overshoot", and "undershoot". Some of these definitions are explained in Figure 1. The numerical method presented here can solve for these quantities without obtaining the full response. Assume that $f(t) \overset{\mathcal{L}}{\leftrightarrow} F(s)$ is a step response of some network. Then

$$f(t) = \frac{-1}{t} \sum_{i=1}^{M} K_i \; F(\frac{z_i}{t})$$

The derivative of f(t) is expressed by

$$f'(t) = \frac{-1}{t} \sum_{i=1}^{M} K_i \; (\frac{z_i}{t}) \; . \; F(\frac{z_i}{t})$$

as discussed above. Newton-Raphson iteration can be used to solve for the time t* such that f(t*) equals a desired value A, (see Figure), by using the iteration scheme

$$t_{k+1} = t_k - \frac{f(t_k)-A}{f'(t_k)} = t_k + \frac{\sum_{i=1}^{M} K_i \, F(z_i/t_k) + A \, t_k}{\sum_{i=1}^{M} K_i \, \frac{z_i}{t_k} \, F(\frac{z_i}{t_k})} \qquad (8)$$

Notice that the derivative is obtained here almost without any additional computational effort, since most of the time will be spent for the evaluation of $F(z_i/t_k)$. Similarly, the time to the overshoot or undershoot can be calculated as the time for which the derivative is equal to zero. The Newton-Raphson formula will then use the first and second derivatives. Tests have shown that the result is usually obtained within five iterations.

All the above discussion assumed that the function F(s) was given in an analytic form. If linear networks are analyzed by a computer, additional information is available and can be used to reduce the error by many orders of magnitude for large values of t. Choose such a step, ΔT, that the method is in the region of accurage results (compare figures in [4]). Calculate the voltages across the capacitors and currents through the inductors at the time ΔT and use these values as new initial conditions of the same network (they can be taken into account as voltage or current sources either in series or in parallel with the energy-storing elements). Another step ΔT can be made and so on. This resetting of the origin of the time axis keeps the numerical inversion in the interval of high numerical accuracy even for large times as demonstrated by a simple example. Figure 2a shows the network considered. The exact impulse response is $f(t) = 1/2e^{-1/2t}$ with $F(s) = 0.5/(s+0.5)$. The resetting is equivalent to

$$f(t) = \mathcal{L}^{-1} \left[\frac{1}{2(s+0.5)} + \frac{C.V_{out}}{s+0.5} \right]$$

The approximation $\xi_{8,10}(z)$ was used with several ΔT. Figure 2b shows the case ΔT=1. It can be seen that for small t the direct inversion is better. From approximately t = 8 the resetting is much better and keeps the relative error below 10^{-10}, many orders of magnitude below the error of the inversion without resetting.

The idea of resetting can also be used for nonlinear networks approximated by piecewise linear segments. For each segment, the network is linear and numerical inversion can be applied. What must be determined is the time at which the calculated f(t) crosses from one linear segment to another. Since the voltage (or current) values of such break points are known from the piecewise linear approximation, the appropriate t_{break} can be found by the Newton-Raphson iteration explained above. The solution of the piecewise-linear networks in the time domain is then essentially equivalent to the search for such t_k which correspond to successive break points. The feasibility of this approach was also tested and found very efficient [7].

References:

1. J. Vlach: Numerical Method for transient responses of linear networks with lumped, distributed or mixed parameters. Journal of the Franklin Institute, Vol. 288, No. 2, August 1969, pp. 99-113.

2. K. Singhal and J. Vlach: Program for numerical inversion of the Laplace transform. Electronics Letters, Vol. 7, No. 14, pp. 413-415, 1971.

3. M. R. Nakhla (Mrs.), K. Singhal, and J. Vlach: Simple method for numerical inversion of the Laplace transform. Proceedings of the Sixteenth Midwest Symposium, University of Waterloo, Waterloo, Ontario, Canada, April 12-13, 1973. Second volume.

4. M. Nakhla, K. Singhal, and J. Vlach: Simple program for numerical inversion of the Laplace transform. Electronics Letters, p.313, 1973.

5. M. Abramowitz and I.A. Stegun: Handbook of Mathematical Functions, Dover Publications, New York, 1965, pp. 255.

6. R. Piessens: On a numerical method for the calculation of transient responses. Journal of the Franklin Institute, Vol. 292, No. 1, July 1971, pp. 57-64.

7. I. Hajj, K. Singhal, J. Vlach, and P.R. Bryant, WATAND - a program for the analysis and design of linear and piecewise linear networks. Proceedings of the Sixteenth Midwest Symposium, University of Waterloo, Waterloo, Ontario, Canada, April 12-13, 1973. First Volume.

Captions to Figures:

Figure 1. Step response and the explanation of some definitions.
Figure 2 a. Network for the example
Figure 2 b. Error curves for direct inversion and inversion by resetting:
  curve 1: relative error for direct inversion
  curve 2: relative error by resetting
  curve 3: absolute error for direct inversion
  curve 4: absolute error by resetting
  curve 5: exact response

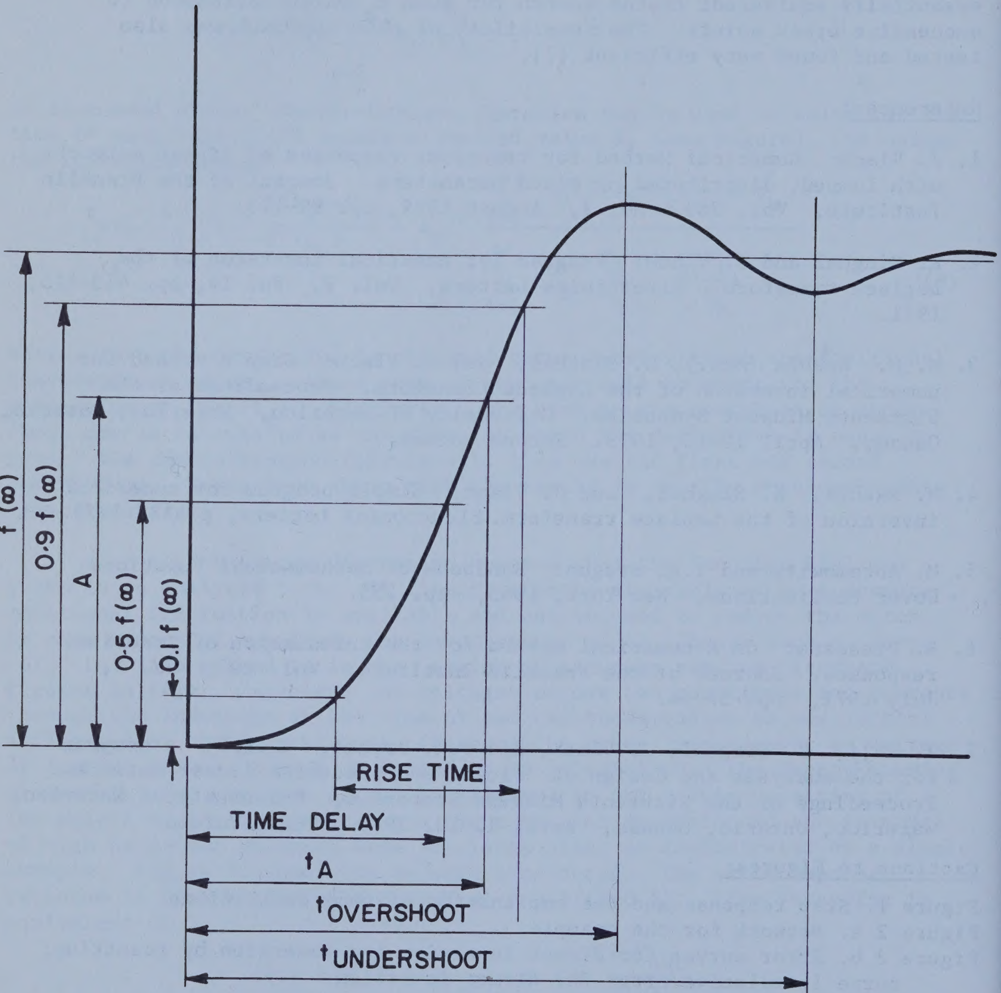

f (∞)
0·9 f (∞)
A
0·5 f (∞)
0.1 f (∞)
RISE TIME
TIME DELAY
t_A
$t_{OVERSHOOT}$
$t_{UNDERSHOOT}$

Figure 1

Figure 2 a

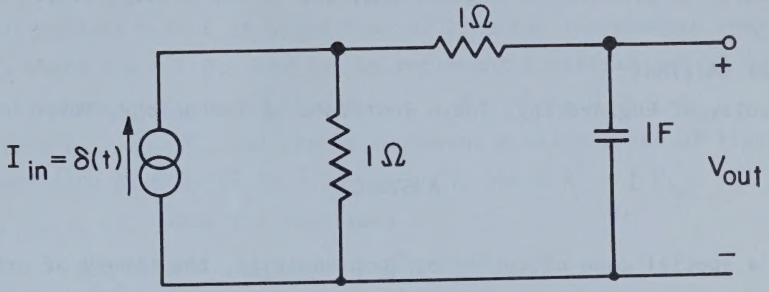
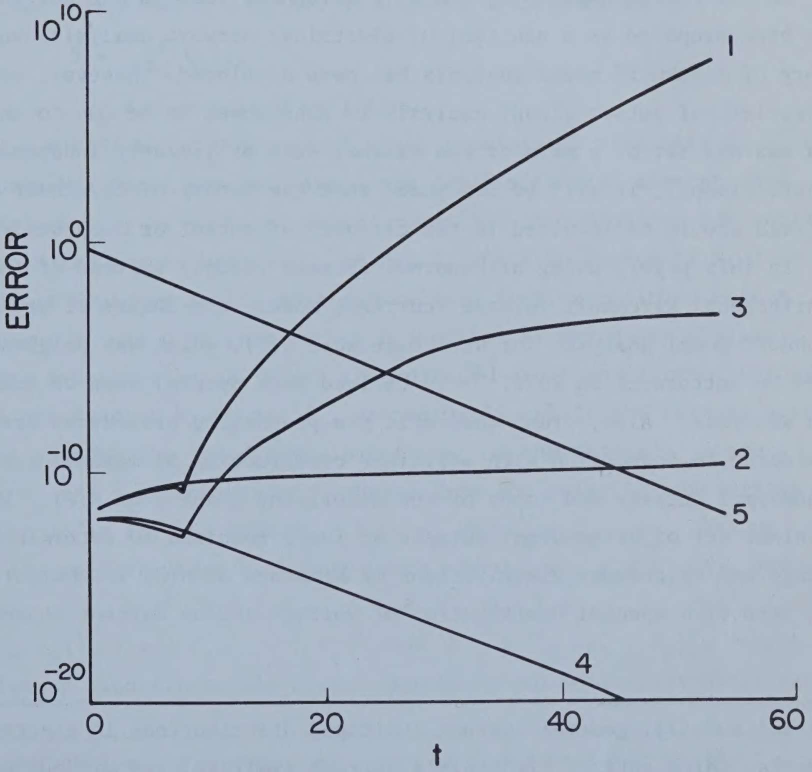

1Ω
$I_{in}=\delta(t)$
1Ω
1F
+
V_{out}
−

Figure 2 b



FURTHER CONSIDERATIONS ON THE THEORY OF ORTHOGONAL KIRCHHOFF MODAL ANALYSIS

SHOJI SHINODA
Faculty of Science and Engineering, Chu-oh University, Tokyo
and
YOJI KAJITANI
Faculty of Engineering, Tokyo Institute of Technology, Tokyo

ABSTRACT

As a special case of cutset or loop analysis, the theory of orthogonal Kirchhoff modal analysis originated by Butterweck in 1972 is derived. Also, two of its related problems are treated.

1. INTRODUCTION

In two recent papers [2] and [3], Kirchhoff voltage and current modes * have been proposed as a new tool of electrical network analysis, and the theory of Kirchhoff modal analysis has been developed. However, so far as the variety of cutset (loop) analysis is considered to be due to the choice of a maximal set or a pair of two maximal sets of linearly independent cutsets (loops), it will be concluded that the theory of Kirchhoff modal analysis should be involved in the category of cutset or loop analysis.

In this paper, using orthonormal cutsets (loops) instead of the terms of orthogonal Kirchhoff voltage (current) modes, the theory of orthogonal Kirchhoff modal analysis for an LCR network N(G), which was originally given by Butterweck in 1972,[2] is derived as a special case of cutset or loop analysis. Also, graph-theoretic pre-processing procedures are considered in connection with efficient construction of maximal sets of orthonormal cutsets and loops of the underlying graph G of N(G). Moreover, a minimal set of orthonormal cutsets or loops required to determine all the voltage and/or current distributions of branches of N(G) is characterized in a case of a special distribution of voltage and/or current sources.

---

* In [2] and [3], general current (voltage) distributions in electrical networks, which satisfy Kirchhoff's current (voltage) law without necessarily satisfying the remaining network constraints, are called the Kirchhoff voltage (current) modes.

2. RECAPITULATIONS OF CUTSET ANALYSIS AND LOOP ANALYSIS FOR AN LCR NETWORK

Let G be a connected graph whose nodes are labelled $1, 2, \ldots, B, B+1$ and whose branches are labelled $1, 2, \ldots, A$. Then the number of linearly independent cutsets * of G is B and that of linearly independent loops ** of G is C, where $C = A - B$. Now let us represent a maximal set of linearly independent cutsets of G by $\{\mathrm{Q}_b \mid b = 1, \ldots, B\}$, where $\mathrm{Q}_b = [\, Q_{1b}, \ldots, Q_{Ab}]^{\mathrm{t}}$ *** $(b = 1, \ldots, B)$, and let us represent a maximal set of linearly independent loops of G by $\{\mathrm{R}_c \mid c = 1, \ldots, C\}$, where $\mathrm{R}_c = [\, R_{1c}, \ldots, R_{Ac}]^{\mathrm{t}}$ $(c = 1, \ldots, C)$. Then the relations

$$\mathrm{Q}_b^{\mathrm{t}}\mathrm{R}_c = 0 \qquad \text{and/or} \qquad \mathrm{R}_c^{\mathrm{t}}\mathrm{Q}_b = 0$$

are satisfied, where $b = 1, \ldots, B$ and $c = 1, \ldots, C$, and if $\mathrm{Q}_b^{\mathrm{t}}\mathrm{I} = 0$, $b = 1, \ldots, B$, then I represents a loop of G and is expressed as

$$\mathrm{I} = \sum_{c=1}^{C} \mathrm{R}_c J_c,$$

where $J_1, \ldots, J_C$ are auxiliary variables, and moreover if $\mathrm{R}_c^{\mathrm{t}}\mathrm{V} = 0$, $c = 1, \ldots, C$, then V represents a cutset of G and is expressed as

$$\mathrm{V} = \sum_{b=1}^{B} \mathrm{Q}_b U_b,$$

where $U_1, \ldots, U_B$ are auxiliary variables.

Next, let N(G) be an LCR network whose topological structure is G, any of whose branches has a non-zero impedance (or admittance) element, and which is driven only by voltage sources connected into branches and current sources connected across branches. In Fig.1, the a-th branch of N(G) $(a = 1, \ldots, A)$ is depicted in its most generalized form, containing both a voltage source e_a and a current source s_a, either of which may obviously be assigned any arbitrary value, where the voltage across and current through the branch are denoted by V_a and I_a, respectively, while the voltage across and current through the impedance element z_{aa} (or admittance y_{aa}) are denoted by v_a and i_a, respectively. Here, from the polarity conventions shown in Fig.1, it follows that $V_a = v_a - e_a$ and $I_a = i_a - s_a$. Let us represent the vectors whose a-th components are v_a, i_a, e_a, s_a, V_a, and I_a by $\mathrm{v} = [\, v_1, \ldots, v_A]^{\mathrm{t}}$, $\mathrm{i} = [\, i_1, \ldots, i_A]^{\mathrm{t}}$, $\mathrm{e} = [\, e_1, \ldots, e_A]^{\mathrm{t}}$, $\mathrm{s} = [\, s_1, \ldots, s_A]^{\mathrm{t}}$, $\mathrm{V} = [\, V_1, \ldots, V_A]^{\mathrm{t}}$, and $\mathrm{I} = [\, I_1, \ldots, I_A]^{\mathrm{t}}$, respectively. Let us also denote the branch impedance and admittance matrices of

---

*, ** The definitions of cutsets and loops of G may be found in [4].

*** The superscript " t " is used to denote the transposition as a matrix.

N(G) by $z = \mathrm{diag}[\, z_{11}, \ldots , z_{AA}]$ and $y = \mathrm{diag}[\, y_{11}, \ldots , y_{AA}]$, respectively, where the symbol " diag[ ] " is used to denote a diagonal matrix (i.e., $\mathrm{diag}[\, y_{11}, \ldots , y_{AA}]$ denotes a matrix whose diagonal elements are $y_{11}, \ldots , y_{AA}$ and whose off-diagonal elements are all zeros.). Here, if we denote a unit matrix of order A by 1, then we have $zy = yz = 1$.

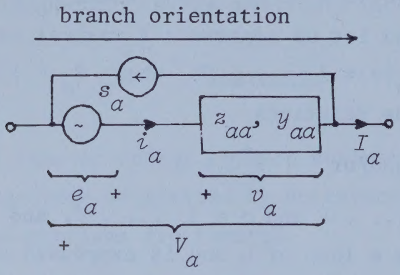


Fig.1 The a-th branch and its associated parameters

CUTSET ANALYSIS FOR N(G)

$Q_b^t i = Q_b^t s,\ b = 1, \ldots , B,$

$v = e + \sum_{b=1}^{B} \alpha Q_b U_b,$

$i = yv,\ V = v - e,\ I = i - s$

$\downarrow$

$\sum_{b=1}^{B} \{Q_{b'}^t (\alpha y) Q_b\} U_b = Q_{b'}^t (s - ye),$

$b' = 1, \ldots , B$

$\downarrow$ if $Y \neq 0$, then U_b, $b = 1, \ldots , B$, are determined and then

$V = \sum_{b=1}^{B} \alpha Q_b U_b$

$\downarrow$

$v = V + e$

$\downarrow$

$i = yv$

$\downarrow$

$I = i - s$

LOOP ANALYSIS FOR N(G)

$R_c^t v = R_c^t e,\ c = 1, \ldots , C,$

$i = s + \sum_{c=1}^{C} \alpha R_c J_c,$

$v = zi,\ I = i - s,\ V = v - e$

$\downarrow$

$\sum_{c=1}^{C} \{R_{c'}^t (\alpha z) R_c\} J_c = R_{c'}^t (e - zs),$

$c' = 1, \ldots , C$

$\downarrow$ if $Z \neq 0$, th n J_c, $c = 1, \ldots , C$, are determined and then

$I = \sum_{c=1}^{C} \alpha R_c J_c$

$\downarrow$

$i = I + s$

$\downarrow$

$v = zi$

$\downarrow$

$V = v - e$

Fig.2 A flow-chart for cutset and loop analyses for N(G) [8]

Thus, under the above preparations, we see the following: Ohm's law can be expressed as $v = zi$ or $i = yv$, and Kirchhoff's current and voltage laws can be expressed as

$$Q_b^t i = Q_b^t s,\ b = 1, \ldots, B, \tag{1}$$

and

$$R_c^t v = R_c^t e,\ c = 1, \ldots, C, \tag{2}$$

respectively, and moreover the solutions of (1) and (2) can be obtained as

$$i = s + \sum_{c=1}^{C} \alpha R_c J_c \qquad \text{and} \qquad v = e + \sum_{b=1}^{B} \alpha Q_b U_b\ ,$$

respectively, where α is assigned any arbitrary non-zero value which will be called the scale factor, $J_1, \ldots, J_C$ are called the loop currents and $U_1, \ldots, U_B$ are called the cutset voltages of N(G).

Hence it follows that both cutset and loop analyses for N(G) are recapitulated in a flow-chart of Fig.2, where Y is the determinant of the $B \times B$ matrix whose b'-th row and b-th column element is $Q_{b'}^t(\alpha y)Q_b$ $(b, b' = 1, \ldots, B)$ and Z is the determinant of the $C \times C$ matrix whose c'-th row and c-th column element is $R_{c'}^t(\alpha z)R_c$ $(c, c' = 1, \ldots, C)$.

3. A DERIVATION OF THE ORTHOGONAL KIRCHHOFF MODAL ANALYSIS AS A CUTSET OR LOOP ANALYSIS

Now let us represent a maximal independent set of elementary cutsets of G by $\{\hat{Q}_b | b = 1, \ldots, B\}$, where $\hat{Q}_b = [\hat{Q}_{1b}, \ldots, \hat{Q}_{Ab}]^t$ $(b = 1, \ldots, B)$, and let us assume that the set $\{Q_b | b = 1, \ldots, B\}$ introduced in the foregoing chapter has been constructed from the set $\{\hat{Q}_b | b = 1, \ldots, B\}$ as follows:

$$Q_1 = \frac{\hat{Q}_1}{\sqrt{\hat{Q}_1^t(\alpha y)\hat{Q}_1}}\ , \quad Q_b = \frac{\tilde{Q}_b}{\sqrt{\tilde{Q}_b^t(\alpha y)\tilde{Q}_b}}\ ,\ b = 2, \ldots, B \tag{3}$$

where

$$\tilde{Q}_b = \hat{Q}_b - \sum_{m=1}^{b-1} \{\hat{Q}_b^t(\alpha y)Q_m\}Q_m\ ,\ b = 2, \ldots, B. \tag{4}$$

Dually, let us also represent a maximal independent set of elementary loops of G by $\{\hat{R}_c | c = 1, \ldots, C\}$, where $\hat{R}_c = [\hat{R}_{1c}, \ldots, \hat{R}_{Ac}]^t$ $(c = 1, \ldots, C)$, and let us assume that the set $\{R_c | c = 1, \ldots, C\}$ introduced in the foregoing chapter has been constructed from the set $\{\hat{R}_c | c = 1, \ldots, C\}$ as follows:

$$R_1 = \frac{\hat{R}_1}{\sqrt{\hat{R}_1^t(\alpha z)\hat{R}_1}}\ , \quad R_c = \frac{\tilde{R}_c}{\sqrt{\tilde{R}_c^t(\alpha z)\tilde{R}_c}}\ ,\ c = 2, \ldots, C \tag{5}$$

where

$$\tilde{R}_c = \hat{R}_c - \sum_{m=1}^{c-1} \{\hat{R}_c^t(\alpha z)R_m\}R_m\ ,\ c = 2, \ldots, C. \tag{6}$$

Then the relations

$$Q^t_{b'}(\alpha y)Q_b = \delta_{b'b},\ R^t_{c'}(\alpha z)R_c = \delta_{c'c},\ Q^t_b R_c = 0, \text{ and } R^t_c Q_b = 0 \qquad (7)$$

are satisfied, where $b, b' = 1, \dots, B$; $c, c' = 1, \dots, C$; and $\delta_{b'b}$ and $\delta_{c'c}$ are used as the Kronecker's deltas. In general, the B column-vectors $Q_1, \dots, Q_B$ satisfying (7) are called the orthonormal cutsets with respect to αy of G and the C column-vectors $R_1, \dots, R_C$ satisfying (7) are called the orthonormal loops with respect to αz of G.

Consequently, if we choose a maximal set of orthonormal cutsets with respect to αy of G as $\{Q_b | b = 1, \dots, B\}$, then the cutset voltages $U_1, \dots, U_B$ are determined simply as

$$U_b = Q^t_b(s - ye),\ b = 1, \dots, B. \qquad (8)$$

In [2], such a cutset analysis has been called the orthogonal Kirchhoff modal analysis for N(G) and the column-vectors $Q_1. \dots, Q_B$ satisfying (7) have been called the orthogonal Kirchhoff voltage modes. Also, if we choose a maximal set of orthonormal loops with respect to αz of G as $\{R_c | c = 1, \dots, C\}$, then the loop currents $J_1, \dots, J_C$ are determined simply as

$$J_c = R^t_c(e - zs),\ c = 1, \dots, C. \qquad (9)$$

In [2], such a loop analysis has been called the orthogonal Kirchhoff modal analysis for N(G) and the column-vectors $R_1, \dots, R_C$ satisfying (7) have been called the orthogonal Kirchhoff current modes.

4. EFFICIENT CONSTRUCTION OF MAXIMAL SETS OF ORTHONORMAL CUTSETS AND LOOPS

In connection with the theory of orthogonal Kirchhoff modal analysis, how to construct either a maximal set of orthonormal cutsets with respect to αy of G, denoted by $\{Q_b | b = 1, \dots, B\}$ hereafter, or a maximal set of orthonormal loops with respect to αz of G, denoted by $\{R_c | c = 1, \dots, C\}$ hereafter, becomes of deep concern to us.

Now, if two elementary cutsets (loops) of G do not have any common branches, then they are said to be combinatorially independent.

If a maximal set of elementary cutsets of G is represented by $\{\hat{Q}_b | b = 1, \dots, B\}$, and if the first F ($\leq B$) vectors $\hat{Q}_1, \dots, \hat{Q}_F$ of $\{\hat{Q}_b | b = 1, \dots, B\}$ are combinatorially independent, then the corresponding vectors $Q_1, \dots, Q_F$ of $\{Q_b | b = 1, \dots, B\}$ can be constructed simply as

$$Q_f = \frac{\hat{Q}_f}{\sqrt{\hat{Q}_f^t(\alpha y)\hat{Q}_f}} \quad , f = 1, \ldots , F$$

without any calculations of (4) for $f = 2, \ldots , F$. From this fact, we see that the larger F is, the more efficient the construction of $\{Q_b | b = 1, \ldots , B\}$ from $\{\hat{Q}_b | b = 1, \ldots , B\}$ is. Accordingly, we see that the problem of finding a maximal set of combinatorially independent cutsets of G is of importance in connection with an efficient construction of $\{Q_b | b = 1, \ldots , B\}$ from G. On the other hand, we also see from the dual treatment that the problem of finding a maximal set of combinatorially independent loops of G is of importance in connection with an efficient construction of $\{R_c | c = 1, \ldots , C\}$ from G. Unfortunately, however, these two problems are left as presently still unsolved problems.*

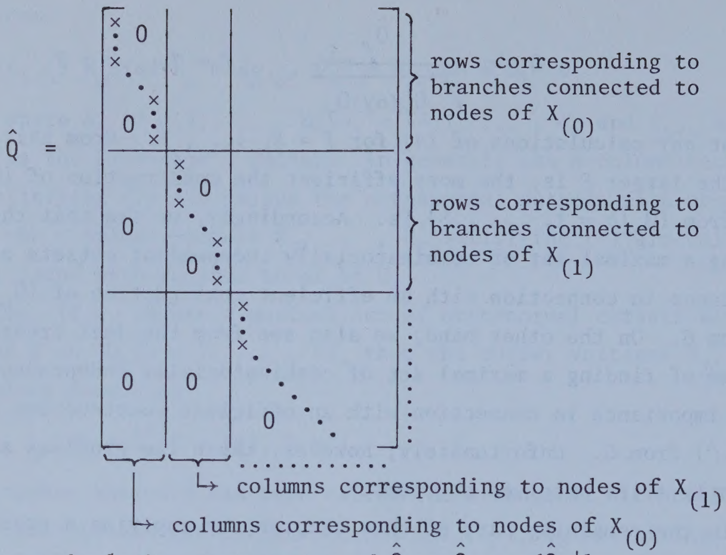
In the remaining part of this chapter, by applying a graph-theoretic pre-processing procedure which has played an important rôle in the applications of sparsity of nodal admittance matrix,[9] an available method is given which constructs $\{Q_b | b = 1, \ldots , B\}$ from the reduced branch-node incidence matrix of G.

Let $G_{(0)} = G$. For each $k = 1, \ldots , K$, let $X_{(k-1)}$ be the maximum internally stable set ** of $G_{(k-1)}$, let $\bar{G}_{(k)}$ be a graph obtained from $G_{(k-1)}$ by adding a new branch between any two nodes of $G_{(k-1)}$ that are both adjacent to each node of $X_{(k-1)}$ but are not adjacent to each other, and let $G_{(k)}$ be a graph obtained from the new graph $\bar{G}_{(k)}$ by removing all nodes of $X_{(k-1)}$ and all branches connected to them of $\bar{G}_{(k)}$, where K is a number k such that the number of nodes of $G_{(k)}$ becomes one. Also, let $\hat{Q}$ be the reduced branch-node incidence matrix of G whose rows correspond to the branches of G and whose columns correspond to the nodes of $X_{(0)}, X_{(1)}, \ldots$, and $X_{(K-1)}$, and let us arrange the columns of $\hat{Q}$ in the order of the nodes of $X_{(0)}, X_{(1)}, \ldots$, and $X_{(K-1)}$, and further let us arrange the rows of $\hat{Q}$ in the order of the branches connected to the nodes of $X_{(0)}, X_{(1)}, \ldots$, and $X_{(K-1)}$. Then $\hat{Q}$ is reduced to the form

---

* However, it should be noted that a characterization of a maximal set of combinatorially independent cutsets of G has already been made by Kishi and Uetake in [6].

** A subset X of the node set of a graph is said to be internally stable if no two nodes of X are adjacent.[1]


$\hat{Q}$ =
×
×
0
rows corresponding to branches connected to nodes of $X_{(0)}$
rows corresponding to branches connected to nodes of $X_{(1)}$
columns corresponding to nodes of $X_{(1)}$
columns corresponding to nodes of $X_{(0)}$

(10)

Here, if we choose the b-th column-vector of $\hat{Q}$ as $\hat{Q}_b$ in $\{\hat{Q}_b | b = 1, \ldots , B\}$, and if we denote the number of nodes in each $X_{(k-1)}$ $(k = 1, \ldots , K)$ by $F(k) - F(k-1)$ where $F(0) = 0$, then the orthonormal cutsets $Q_1, \ldots , Q_{F(1)}$ corresponding to $\hat{Q}_1, \ldots , \hat{Q}_{F(1)}$, respectively, can be determined as

$$Q_{f(1)} = \frac{\hat{Q}_{f(1)}}{\sqrt{\hat{Q}^t_{f(1)}(\alpha y)\hat{Q}_{f(1)}}} , \; f(1) = 1, \ldots , F(1)$$

and, in general, for $k = 2, \ldots , K$, the orthonormal cutsets $Q_{F(k-1)+1}, \ldots , Q_{F(k)}$ corresponding to $\hat{Q}_{F(k-1)+1}, \ldots , \hat{Q}_{F(k)}$, respectively, can be determined as

$$Q_{f(k)} = \frac{\tilde{Q}_{f(k)}}{\sqrt{\tilde{Q}^t_{f(k)}(\alpha z)\tilde{Q}_{f(k)}}} , \; f(k) = F(k-1)+1, \ldots , F(k)$$

where

$$\tilde{Q}_{f(k)} = \hat{Q}_{f(k)} - \sum_{f=1}^{F(k-1)} \{\hat{Q}^t_{f(k)}(\alpha y)Q_f\}Q_f .$$

Thus, this method enables us to construct $\{Q_b | b = 1, \ldots , B\}$ efficiently from G. The desirable features of the method are the following: (i) $\{\hat{Q}_b | b = 1, \ldots , B\}$ can be easily obtained from G; (ii) in construction of $\{Q_{f(1)} | f(1) = 1, \ldots , F(1)\}$ from $\{\hat{Q}_{f(1)} | f(1) = 1, \ldots , F(1)\}$, any calculations of (4) are not required; and (iii) for each $k = 2, \ldots , K$, the respective calculations of $\tilde{Q}_{F(k-1)+1}, \ldots$, and $\tilde{Q}_{F(k)}$ from $\hat{Q}_{F(k-1)+1}, \ldots$, and $\hat{Q}_{F(k)}$ are mutually independent. Here, it should be especially noted that the features (ii) and (iii) follow readily from the distribution of zeros in $\hat{Q}$ which is characterized by the orderings of the columns and the rows of $\hat{Q}$.

5. A MINIMAL SET OF ORTHONORMAL CUTSETS OR LOOPS REQUIRED TO DETERMINE v, i, V and I of N(G)

In the theory of orthogonal Kirchhoff modal analysis for N(G), either $\{Q_b | b = 1, \ldots, B\}$ or $\{R_c | c = 1, \ldots, C\}$ must be constructed, but the number of orthonormal cutsets or loops required to determine v, i, V and I of N(G) is not more than the smallest of B and C.

Let W be a maximal set of branches corresponding to non-zero components of $s - ye$, and let W be the number of elements of W. Then, since z and y are diagonal matrices, and since $e - zs = -z(s - ye)$, W is also the maximal set of branches corresponding to non-zero components of $e - zs$.

Now let us consider the case where W contains neither cutsets nor loops of G. Then there exists at least one pair of tree and cotree of G which contain W in common. Therefore let T be a tree of G which contains W and let $\bar{T}$ be the cotree of G corresponding to T and further let us assume that the branches of $\bar{T}$ have been labelled $1, 2, \ldots, C$, those of T - W have been labelled $C+1, \ldots, A-W$ and those of W have been labelled $A-W+1, \ldots, A$, where T - W is a set of branches of T which are not in W. Then, the fundamental cutset matrix with respect to T of G, denoted by $\hat{Q}$, has the form

$$\hat{Q} = \begin{array}{r|c|c|l}
 & 1 \ \cdots \ B-W & B-W+1 \ \cdots \ B & \\
\hline
1 & & & \\
\vdots & & & \Big\}\ \bar{T} \\
C & & & \\
\hline
C+1 & 1 \qquad 0 & & \\
\vdots & \ddots & 0 & \Big\}\ T-W \\
A-W & 0 \qquad 1 & & \\
\hline
A-W+1 & & 1 \qquad 0 & \\
\vdots & 0 & \ddots & \Big\}\ W \\
A & & 0 \qquad 1 &
\end{array} \tag{11}$$

Here, if the b-th column-vector of $\hat{Q}$ is denoted by $\hat{Q}_b$ in $\{\hat{Q}_b | b = 1, \ldots, B\}$, then, according to (3) and (4), the corresponding set $\{Q_b | b = 1, \ldots, B\}$ can be constructed in such a way that an $A \times B$ matrix Q, whose b-th column-vector is Q_b, has the form

$$
Q = \begin{array}{r|ccc|ccc|l}
 & 1 & \cdots & B\text{-}W & B\text{-}W+1 & \cdots & B & \\
\hline
1 & & & & & & & \\
\vdots & & & & & & & \left.\right\} \bar{T} \\
C & & & & & & & \\
\hline
C+1 & \times & & & & & & \\
\vdots & & \ddots & & & & & \left.\right\} T - W \\
A\text{-}W & & 0 & \times & & & & \\
\hline
A\text{-}W+1 & & & & \times & & & \\
\vdots & & 0 & & & \ddots & & \left.\right\} W \\
A & & & & & 0 & \times & \\
\hline
\end{array}
\tag{12}
$$

Accordingly, U_b, $b = 1, \ldots, B$, can be obtained as

$$U_b = 0, \text{ for } b = 1, \ldots, B\text{-}W; \text{ and} \tag{13}$$

$$U_b = Q_b^t(s - ye) = \sum_{a=A-W+1}^{A} Q_{ab}(s_a - y_a e_a), \text{ for } b = B\text{-}W+1, \ldots, B. \tag{14}$$

In particular, if T = W, then U_b, $b = 1, \ldots, B$, can be obtained as

$$U_b = Q_b^t(s - ye) = \sum_{a=C+1}^{A} Q_{ab}(s_a - y_a e_a), \text{ for } b = 1, \ldots, B. \tag{15}$$

Also, it follows that V, v, i, and I of N(G) can be obtained respectively as $V = \sum_{b=B-W+1}^{B} \alpha Q_b U_b$, $v = V + e$, $i = yv$, and $I = i - s$.

Consequently, from the above and its dual treatment, it is readily seen that, in the case where W contains neither cutsets nor loops of G, the orthonormal cutsets or loops required to determine v, i, V and I of N(G) without any use of the guage and dual guage transformations of sources * are W in number, where W is not more than the smallest of B and C.

On the other hand, in the case where W contains cutsets or loops of G and/or in the case where any use of the guage and dual guage transformations of sources is allowed, minimal sets of orthonormal cutsets or loops required to determine v, i, V and I of N(G) will not be so simply characterized as the above. Because of the confinement of this paper, the characterizations will be prolonged in a future paper.

---

* Let $\{\bar{R}_c \mid c = 1, \ldots, C\}$ be a maximal set of linearly independent loops of G. Then, $\bar{s} = s + \sum_{c=1}^{C} \bar{R}_c \xi_c$, where $\xi_1, \ldots, \xi_C$ are assigned arbitrary values, is called the guage transformation of current sources. Also, let $\{\bar{Q}_b \mid b = 1, \ldots, B\}$ be a maximal set of linearly independent cutsets of G. Then, $\bar{e} = e + \sum_{b=1}^{B} \bar{Q}_b \eta_b$, where $\eta_1, \ldots, \eta_B$ are assigned arbitrary values, is called the dual guage transformation of voltage sources. It should be especially noted that v and i of N(G) will remain unchanged under the guage and dual guage transformations of sources.[7]

6. CONCLUSIONS

It has been shown that the theory of orthogonal Kirchhoff modal analysis for an LCR network belongs to the category of cutset or loop analysis for the network. Also, in connection with the analysis, two problems have been considered and partial solutions have been given to the problems.

7. ACKNOWLEDGMENT

The authers wish to express their thanks to Prof. G. Kishi of Tokyo Institute of Technology, Tokyo, and Prof. H. Ohrui of Chu-oh University, Tokyo, for their continued encouragements.

REFERENCES

[1] C. Berge, "The Theory of Graphs and Its Applications," John Wiley, New York, 1962.

[2] H. J. Butterweck, "Orthogonal Kirchhoff Modes — a New Tool of Network Analysis," IEEE Trans. Circuit Theory, Vol.CT-19, pp.309-312, 1972.

[3] H. J. Butterweck and R. Wilke, "Some Results on Kirchhoff Modes in Electrical Networks," Arch. Elek. Übertragung, Vol.27, pp.197-209, 1973.

[4] M. Iri, "Network Flow, Transportation and Scheduling — Theory and Algorithms," Academic Press, New York, 1969.

[5] G. Kishi and Y. Kajitani, "Topological Considerations on Minimal Sets of Variables Describing an LCR Network," IEEE Trans. Circuit Theory, Vol.CT-20, pp.335-340, 1973.

[6] G. Kishi and Y. Uetake, "On Sparsity of Admittance Matrices," Trans. of IECE of Japan, Vol.55-A, pp.205-212, 1972.

[7] S. Okada and R. Onodera, "Theory of Interlinked Electromagnetic Networks and Fields in Tensor Geometry of Linear Space," Memoirs of RAAG, Vol.1, pp.1-67, 1955.

[8] S. Seshu and M. B. Reed, "Linear Graphs and Electrical Networks," Addison-Wesley, Reading, 1961.

[9] S. Shinoda and Y. Kajitani, "Sparsity of Nodal Admittance Matrix and Its Applications," Trans. of IECE of Japan, Vol.56-A, pp.544-550, 1973.

EXAMPLES ILLUSTRATING THE CONNECTION BETWEEN SINGULAR PERTURBATIONS AND SINGULAR ARCS*

R. E. O'MALLEY, JR.
Department of Mathematics
University of Arizona
Tucson, Arizona 85721

ABSTRACT

Singular perturbation techniques are used to obtain the asymptotic solution of some nearly singular optimal control problems. The resulting outer solutions agree asymptotically with the familiar singular arc solutions.

GENERAL DISCUSSION

The linear regulator problem

$$\left.\begin{array}{l} \dot{x} = Ax + Bu, \quad 0 \le t \le 1 \\ \text{with} \\ x(0) \text{ given} \\ \text{and with the scalar cost} \\ J(\varepsilon) = \frac{1}{2}\int_0^1 (x^T(t)Q(t)x(t) + u^T(t)R(t,\varepsilon)u(t))\,dt \\ \text{to be minimized} \end{array}\right\} \quad (1.1)$$

for x and u vectors of dimension n and r, respectively, and Q and R symmetric positive semi-definite matrices is a well-studied problem in the optimal control literature (cf. e.g., Anderson and Moore (1971)). We shall take $R(t,\varepsilon)$ to be positive definite for $\varepsilon > 0$ and singular for $\varepsilon = 0$. For each fixed $\varepsilon > 0$, a unique optimal control $u(t,\varepsilon)$ is readily obtained (cf. Kalman (1960)), while singular arcs occur when $\varepsilon = 0$ and impulsive controls must be expected (cf. Bryson and Ho (1969)).

A special case where

$$R(t,\varepsilon) = \varepsilon^2 R_o \quad \text{and} \quad Q(t) = Q_o \qquad (1.2)$$

for constant matrices R_o and Q_o is particularly instructive. It is an example of cheap control since the cost of the control u is cheap relative to that of the state x (provided Q is nontrivial) [cf. Lions (1973) who coined the term for an analogous problem for distributed parameter systems]. Such problems have also occurred in studying the limiting possibilities for regulators (cf. Kwakernaak and Sivan (1972)) and inverse problems (cf. Kreindler and Jameson (1974)). They were alluded to in an earlier Allerton session on singular perturbations (cf. Edelbaum (1970)) and have been recently studied by Jameson and the author (cf. Jameson and O'Malley (   ) and Jameson's article in these proceedings).

---

*This research was supported in part at the University of Arizona by the Office of Naval Research.

If we introduce the Hamiltonian

$$H(x,u,p,t,\varepsilon) = \frac{1}{2}(x^TQ(t)x + u^TR(t,\varepsilon)u) + p^T(Ax + Bu)$$

where the costate p satisfies

$$\dot{p} = -\frac{\partial H}{\partial x} = -Qx - A^Tp, \quad p(1) = 0, \tag{1.3}$$

we will have

$$\frac{\partial H}{\partial u} = Ru + B^Tp = 0 \tag{1.4}$$

along an optimal control. This implies the control relation

$$u = -R^{-1}B^Tp \tag{1.5}$$

for $\varepsilon > 0$, while for $\varepsilon = 0$ the control is not immediately determined (though a singular arc condition results).

For $\varepsilon > 0$, then, the control problem has been reduced to the linear boundary value problem

$$\left.\begin{aligned} &\dot{x} = Ax - BR^{-1}B^Tp, \quad x(0) \text{ prescribed} \\ &\dot{p} = -Qx - A^Tp, \quad p(1) = 0 \end{aligned}\right\} \tag{1.6}$$

which is uniquely solvable (cf. Bucy (1967)). We are, however, interested in the asymptotic solution as $\varepsilon \to 0$. To obtain it, we must specify the limiting behavior of $R(t,\varepsilon)$. In the special case where (1.2) holds, (1.6) becomes the singularly perturbed problem

$$\begin{aligned} &\varepsilon^2\dot{x} = \varepsilon^2Ax - BR_o^{-1}B^Tp, \quad x(0) \text{ prescribed} \\ &\dot{p} = -Q_ox - A^Tp, \quad p(1) = 0 \end{aligned} \tag{1.7}$$

and the optimal control is given by

$$u = -\frac{1}{\varepsilon^2}R_o^{-1}B^Tp. \tag{1.8}$$

We note, in particular, that the differential order of the system (1.7) drops from $2n$ when $\varepsilon > 0$ to n when $\varepsilon = 0$. Application of standard singular perturbation theory results in different asymptotic solutions in different cases. Thus Jameson and O'Malley discuss Case 0, where B is of rank n and Q_o is positive definite, and the Cases k, $k \geq 1$, where B has rank $r \leq n$ and k is the least integer such that $B^T(A^T)^{k-1}Q_oA^{k-1}B$ is positive definite. In all cases, initial behavior is impulse-like while behavior within $(0,1)$ follows a singular arc, i.e. $B^Tp = 0$. In Case k, the limiting control near $t = 0$ behaves like the limit of $\frac{1}{\varepsilon}e^{-t/\varepsilon^{1/k}}$ as $\varepsilon \to 0$. We note, in particular, that $\varepsilon^{1/k}$, the boundary layer thickness, increases with k and the number of constraints $B^TQ_oB = B^TA^TQ_oAB = \cdots = B^T(A^T)^{k-2}Q_oA^{k-2}B = 0$, $B^T(A^T)^{k-1}Q_oA^{k-1}B > 0$. Further, the resulting trajectory in Case k has a large initial excursion in the controllability space spanned by $B, AB, \cdots, A^{k-1}B$. Heuristic arguments supporting these singular perturbation results are also given by Jameson and O'Malley.

We will now illustrate these possibilities and some more general problems by examples.

EXAMPLES

1. The scalar problem

$$\left.\begin{array}{l} \dot{x}(t) = u(t), \quad x(0) \text{ given} \\ \text{with} \\ J(\varepsilon) = \frac{1}{2}\int_o^1 (x^2(t) + \varepsilon^2u^2(t))\, dt \end{array}\right\}$$

can be solved by using the Hamilton-Jacobi equations

$$\left.\begin{array}{l} \dot{x} = u, \quad x(0) \text{ given} \\ \dot{p} = -x, \quad p(1) = 0 \end{array}\right\}$$

and the control relation

$$u = -p/\varepsilon^2.$$

Solving this boundary value problem we find the optimal control

$$u(t,\varepsilon) = -\frac{1}{\varepsilon}(e^{-t/\varepsilon} - e^{-1/\varepsilon}e^{-(1-t)/\varepsilon})\left(\frac{x(0)}{1+e^{-2/\varepsilon}}\right)$$

and the corresponding trajectory

$$x(t,\varepsilon) = (e^{-t/\varepsilon} + e^{-1/\varepsilon}e^{-(1-t)/\varepsilon})\left(\frac{x(0)}{1+e^{-2/\varepsilon}}\right)$$

or, asymptotically,

$$(x(t,\varepsilon), u(t,\varepsilon)) \sim (e^{-t/\varepsilon}x(0), -\frac{1}{\varepsilon}e^{-t/\varepsilon}x(0)).$$

Thus the solution features nonuniform convergence at $t = 0$ as $\varepsilon \to 0$ unless $x(0) = 0$ (like the boundary layer function $e^{-t/\varepsilon}$). Away from $t = 0$, the optimal control and the resulting trajectory are asymptotically negligible, as is the trivial singular arc solution appropriate for $\varepsilon = 0$. We finally note that even though the optimal control is initially unbounded, the corresponding limiting cost

$$\int_o^1 e^{-\frac{2t}{\varepsilon}}x^2(0)\, dt = O(\varepsilon)$$

is, indeed, cheap.

2. The regulator problem for a harmonic oscillator might be described by

$$\left.\begin{aligned}&\ddot{y} + y = u, \quad y(0),\ \dot{y}(0) \text{ prescribed}\\ &\text{with}\\ &J(\varepsilon) = \frac{1}{2}\int_0^1 (\dot{y}^2(t) + \varepsilon^2 u^2(t))\,dt.\end{aligned}\right\}$$

Introducing $x_1 = y$ and $x_2 = \dot{y}$, the Hamilton-Jacobi equations imply

$$\left.\begin{aligned}\dot{x}_1 &= x_2 \quad\ \ , & x_1(0) &= y(0)\\ \dot{x}_2 &= -x_1 + u\ , & x_2(0) &= \dot{y}(0)\\ \dot{p}_1 &= p_2 \quad\ \ , & p_1(1) &= 0\\ \dot{p}_2 &= -x_2 - p_1, & p_2(1) &= 0\end{aligned}\right\}$$

with the optimal control given by

$$u = -p_2/\varepsilon^2, \quad \varepsilon > 0.$$

Solving this boundary value problem and expanding the exact solution asymptotically results in an optimal control of the form

$$u(t,\varepsilon) = \left(y(0) - \frac{y'(0)}{\varepsilon} e^{-t/\varepsilon} - \frac{y(0)}{4} e^{(t-1)/\varepsilon}\right)(1 + 0(\varepsilon))$$

with the corresponding trajectories given by

$$x_1(t,\varepsilon) = \left(y(0) - \varepsilon y'(0) e^{-t/\varepsilon} - \varepsilon^2 \frac{y(0)}{4} e^{(t-1)/\varepsilon}\right)(1 + 0(\varepsilon))$$

and

$$x_2(t,\varepsilon) = \left(y'(0) e^{-t/\varepsilon} - \varepsilon \frac{y(0)}{4} e^{(t-1)/\varepsilon}\right)(1 + 0(\varepsilon)) + 0(\varepsilon^2).$$

Note, in particular, that the optimal control behaves like the impulse function $-\frac{y'(0)}{\varepsilon} e^{-t/\varepsilon}$ near $t = 0$. Within $(0,1)$, however, the limiting solution

$$(x_1(t,0), x_2(t,0), u(t,0)) = (y(0), 0, y(0))$$

follows the singular arc, i.e. $B^T p = p_2 = 0$, and satisfies the limiting system

$$\left.\begin{aligned}\dot{x}_1 &= x_2, \quad x_1(0) = y(0)\\ \dot{x}_2 &= -x_1 + u\\ \dot{p}_1 &= 0, \quad p_1(1) = 0\\ 0 &= -x_2\ .\end{aligned}\right\}$$

Also observe that the optimal cost is asymptotic to the trivial cost obtained when following the singular arc. Finally, as Example 4 indicates, quite different limiting behavior would result if the cost functional was changed to

$$J(\varepsilon) = \frac{1}{2}\int_0^1 (y^2(t) + \varepsilon^2 u^2(t))\, dt.$$

3. Now let us consider the first example with the scalar cost restricted such that $|u| \leq 1$. Recall that the optimal control without this constraint was initially unbounded (unless $x(0) = 0$ when the solution was trivial). The solution will again satisfy the Hamilton-Jacobi equations

$$\left.\begin{array}{l} \dot{x} = u, \quad x(0) \ \text{given} \\ \dot{p} = -x, \ p(1) = 0 \end{array}\right\}$$

but the optimal control will now be given by

$$u = -\text{sat}(p/\varepsilon^2) = \begin{cases} -1 & \text{if} \ \ p > \varepsilon^2 \\ -p/\varepsilon^2 & \text{if} \ \ |p| \leq \varepsilon^2 \\ 1 & \text{if} \ \ p < \varepsilon^2. \end{cases}$$

Clearly, if the zero state is reached, the solution will thereafter be trivial. For $x(0) \neq 0$, then, the solution of Example 1 suggests that $|x|$ should be nonincreasing and u should begin as

$$u = -\text{sgn}(x(0)) = \begin{cases} -1 & \text{if} \ \ x(0) > 0 \\ \\ 1 & \text{if} \ \ x(0) < 0. \end{cases}$$

This should continue as $|p|$ decreases monotonically to ε^2. Then, at $t = \tilde{t}$, say, we will switch to the control $u = -p/\varepsilon^2$ until we attain $p(1) = 0$.

We can construct this solution from the differential equations. Thus, for $0 \leq t \leq \tilde{t}$,

$$\left.\begin{array}{l} u(t,\varepsilon) = -\text{sgn}(x(0)) \\ x(t,\varepsilon) = \text{sgn}(x(0))(|x(0)| - t) \\ \text{and} \\ p(t,\varepsilon) = \text{sgn}(x(0))(\frac{t^2}{2} - |x(0)|t + |p(0)|). \end{array}\right\}$$

Since $\tilde{t}$ is the least time such that $|p(\tilde{t})| = \varepsilon^2$, we have

$$\tilde{t} = |x(0)|[1 - \sqrt{1 - \frac{2}{x^2(0)}(|p(0)| - \varepsilon^2)}].$$

Then, for $\tilde{t} \leq t \leq 1$, $u = -p/\varepsilon^2$ so integration implies

$$x(t,\varepsilon) = c_1 e^{-(t-\tilde{t})/\varepsilon} + c_2 e^{-(1-t)/\varepsilon}$$

and

$$p(t,\varepsilon) = \varepsilon[c_1 e^{-(t-\tilde{t})/\varepsilon} - c_2 e^{-(1-t)/\varepsilon}].$$

The condition $p(1) = 0$ implies that $c_2 = c_1 e^{-(1-\tilde{t})/\varepsilon}$ while

$p(\tilde{t}) = \operatorname{sgn}(x(0))\varepsilon^2$ implies that $c_1 = \dfrac{\varepsilon \operatorname{sgn}(x(0))}{1 - e^{-2(1-\tilde{t})/\varepsilon}}$. Thus, we have

$$x(t,\varepsilon) = \varepsilon\left(e^{-(t-\tilde{t})/\varepsilon} + e^{-(1-\tilde{t})/\varepsilon}e^{-(1-t)/\varepsilon}\right)\left(\frac{\operatorname{sgn}(x(0))}{1 - e^{-2(1-\tilde{t})/\varepsilon}}\right)$$

$$u(t,\varepsilon) = -\left(e^{-(t-\tilde{t})/\varepsilon} - e^{-(1-\tilde{t})/\varepsilon}e^{-(1-t)/\varepsilon}\right)\left(\frac{\operatorname{sgn}(x(0))}{1 - e^{-2(1-\tilde{t})/\varepsilon}}\right)$$

for $t \geq \tilde{t}$. Finally, for x to be continuous at $\tilde{t}$, we have

$$\frac{\varepsilon \operatorname{sgn}(x(0))}{(1 - e^{-2(1-\tilde{t})/\varepsilon})}(1 + e^{-2(1-\tilde{t})/\varepsilon}) = \operatorname{sgn}(x(0))(|x(0)| - \tilde{t}).$$

Introducing $\zeta = \dfrac{1-\tilde{t}}{\varepsilon}$, $\tilde{t}$ will be determined by the largest root ζ of the transcendental equation

$$\frac{1 + e^{-2\zeta}}{1 - e^{-2\zeta}} = \frac{|x(0)| - 1}{\varepsilon} + \zeta,$$

i.e.

$$f(\zeta) \equiv \coth \zeta - \zeta = (|x(0)| - 1)/\varepsilon.$$

Since $f'(\zeta) < 0$, f decreases monotonically from $f(0) = \infty$ to $f(\infty) = -\infty$. Thus we have a unique root ζ and a unique switching time $\tilde{t}$. In particular, if $|x(0)| = 1$, the root would be $\zeta \approx 1.2$, so $\tilde{t} \approx 1-1.2\varepsilon$. If $0 < |x(0)| < 1$, we'd have $\zeta \to \infty$ and $\tilde{t} \approx |x(0)| - \varepsilon$, while if $|x(0)| > 1$, $\zeta \to 0$ and $\tilde{t} \approx 1 - \varepsilon^2/(|x(0)| - 1)$.

In summary, then, the limiting optimal control as $\varepsilon \to 0$ is given by

$$u(t) = \begin{cases} -\operatorname{sgn}(x(0)), & 0 \leq t \leq \hat{t} \equiv \min(1, |x(0)|) \\ 0, & \hat{t} \leq t \leq 1. \end{cases}$$

The control lies at the appropriate endpoint of the constraint set until $t = \hat{t} \approx \tilde{t}$, at which time boundary layer behavior occurs. Beyond $\hat{t}$, the trivial singular arc solution will be followed. (Lions has showed us that analogous results can be obtained by use of variational inequalities.)

4. The harmonic oscillator problem

$$\left.\begin{aligned} &\ddot{y} + y = u_1 + u_2,\ y(0),\ \dot{y}(0) \text{ prescribed} \\ &\text{with} \\ &J(\varepsilon,\mu) = \frac{1}{2}\int_0^1 (y^2 + \varepsilon^2 u_1^2 + \varepsilon^2\mu^2 u_2^2)\,dt \\ &\text{to be minimized} \end{aligned}\right\}$$

for ε and μ both small positive parameters is an example of a regulator problem with two cheap controls u_1 and u_2, but with u_2 cheaper than u_1. Introducing $x_1 = y$, $x_2 = \dot{y}$, and the Hamiltonian

$$H(x,u,p,\varepsilon) = \frac{1}{2}\,(x_1^2 + \varepsilon^2 u_1^2 + \varepsilon^2 \mu^2 u_2^2) + p_1 x_2 + p_2(-x_1 + u_1 + u_2),$$

the Hamilton-Jacobi equations become

$$\left.\begin{array}{ll} \dot{x}_1 = x_2 & , \quad x_1(0) = y(0) \\ \dot{x}_2 = -x_1 + u_1 + u_2, & x_2(0) = \dot{y}(0) \\ \dot{p}_1 = -x_1 + p_2 & , \quad p_1(1) = 0 \\ \dot{p}_2 = -p_1 & , \quad p_2(1) = 0 \end{array}\right\}$$

while optimality implies

$$\frac{\partial H}{\partial u_1} = \varepsilon^2 u_1 + p_2 = 0 \quad \text{and} \quad \frac{\partial H}{\partial u_2} = \varepsilon^2 \mu^2 u_2 + p_2 = 0.$$

Substituting for u_1 and u_2 in the preceding system results in a singularly perturbed boundary value problem that can be readily integrated. In particular, x_1 will satisfy the fourth-order equation

$$\varepsilon^2 \mu^2 x_1^{(4)} + 2\varepsilon^2 \mu^2 x_1^{(2)} + (1 + \mu^2 + \varepsilon^2 \mu^2) x_1 = 0.$$

Applying the boundary conditions to its general solution implies that

$$x_1(t,\varepsilon,\mu) = \mathrm{Re}\left[y(0)(1+i)\rho_2(\sqrt{\varepsilon\mu},\mu^2) \exp\left(\frac{-e^{\frac{i\pi}{4}} \rho_1(\sqrt{\varepsilon\mu},\mu^2) t}{\sqrt{\varepsilon\mu}} \right) \right]$$

where ρ_1 and ρ_2 have double asymptotic series expansion in $\sqrt{\varepsilon\mu}$ and μ^2 with leading coefficient equal to one. In particular, we note that $y = x_1$ decays to zero like $e^{-t/\sqrt{2\varepsilon\mu}}$. The optimal controls are given by

$$u_1(t,\varepsilon,\mu) = \frac{\mu}{\varepsilon(1+\mu^2)}\,\mathrm{Re}\left[y(0)(1-i)\rho_3(\sqrt{\varepsilon\mu},\mu^2) \exp\left(\frac{-e^{\frac{i\pi}{4}} \rho_1(\sqrt{\varepsilon\mu},\mu^2) t}{\sqrt{\varepsilon\mu}} \right) \right]$$

and

$$u_2(t,\varepsilon,\mu) = \frac{1}{\varepsilon\mu(1+\mu^2)}\,\mathrm{Re}\left[y(0)(1-i)\rho_3(\sqrt{\varepsilon\mu},\mu^2) \exp\left(\frac{-e^{\frac{i\pi}{4}} \rho_1(\sqrt{\varepsilon\mu},\mu^2) t}{\sqrt{\varepsilon\mu}} \right) \right]$$

where $\rho_3 = 1 + O(\sqrt{\varepsilon\mu}) + O(\mu^2)$ has a double asymptotic series expansion.

Thus, u_2 is initially unbounded like the impulse $\frac{1}{\varepsilon\mu} e^{-t/\sqrt{2\varepsilon\mu}} y(0)$, while $u_1(0,\varepsilon,\mu)$ is bounded or unbounded depending on the ratio $\frac{\mu}{\varepsilon}$. In either case, both controls decay to the trivial solution within $(0,1)$ (corresponsing to the singular arc solution for $\varepsilon = \mu = 0$), while the cheaper control u_2 has a much larger magnitude than u_1 in the initial boundary layer.

REFERENCES

1. B.D.O. Anderson and J.B. Moore, Linear Optimal Control, Prentice-Hall, Englewood-Cliffs, New Jersey, 1971.
2. A.E. Bryson and Y-C Ho, Applied Optimal Control, Blaisdell, Waltham, Massachusetts, 1969.
3. R.S. Bucy, "Two-point boundary value problems of linear Hamiltonian systems," SIAM J. Applied Math. 15 (1967), pp. 1385-1389.
4. T.N. Edelbaum, "Singular perturbations and minimal fuel space trajectories," Proceedings, Eighth Annual Allerton Conf. Circuit and System Theory, University of Illinois, Urbana, 1970, pp. 348-356.
5. A. Jameson and R.E. O'Malley, Jr., "Singular perturbations and singular arcs," to appear.
6. R.E. Kalman, "Contributions to the theory of optimal control," Bol. Soc. Mat. Mexicana 5 (1960), pp. 102-109.
7. E. Kreindler and A. Jameson, to appear.
8. H. Kwakernaak and R. Sivan, "The maximally achievable accuracy of linear optimal regulators and linear optimal filters," IEEE Trans. Automatic Control 17 (1972), pp. 79-86.
9. J.L. Lions, Perturbations Singulières dans les Problèmes aux Limites en Contrôle Optimal, Lecture Notes in Mathematics 323, Springer-Verlag, Berlin, 1973.

THE CONNECTION BETWEEN SINGULAR PERTURBATIONS AND SINGULAR ARCS: PART 2: A THEORY FOR THE LINEAR REGULATOR *

Antony Jameson
Courant Institute
New York University
New York, New York 10012

ABSTRACT

Singular perturbation theory is applied to obtain the asymptotic solution for nearly singular optimal control of a constant linear system in a finite time interval. In the limit as the control cost is reduced to zero, the initial control is found to have an impulse-like behavior, while the outer solution yields a singular arc.

INTRODUCTION

O'Malley's paper at these proceedings discusses several examples of nearly singular optimal control. The cases where all components of the control are equally 'cheap', and unbounded, can be incorporated within the framework of the theory for the constant linear regulator sketched in the following paragraphs. A more detailed treatment, using a preliminary transformation to canonical form, is given in a forthcoming paper (O'Malley and Jameson, 1974).

Consider the constant linear system

$$\dot{x} = Ax + Bu \ , \quad x(0) \text{ specified}, \tag{1}$$

where the n dimensional vector x represents the state, and the r dimensional vector u is the control. Let u be chosen to minimize

$$J = \frac{1}{2} \int_0^1 (x^T Q x + \varepsilon^2 u^T R u)\, dt$$

where Q is a nonnegative definite constant matrix, and R is a positive definite constant matrix. Then

$$u = -\frac{1}{\varepsilon^2} R^{-1} B^T p \tag{3}$$

where x and p satisfy the Hamiltonian equations (Bryson and Ho, 1969)

$$\left.\begin{aligned} \varepsilon^2 \dot{x} &= \varepsilon^2 A x - B R^{-1} B^T p \ , & x(0) \text{ specified} \\ \dot{p} &= -Qx - A^T p \ , & p(1) = 0 \ . \end{aligned}\right\} \tag{4}$$

We shall examine the asymptotic behavior of the solution as $\varepsilon \to 0$. The system of equations (4) is singularly perturbed because its order is reduced from 2n to n when $\varepsilon = 0$. It proves convenient to distinguish the following hierarchy of cases:

---

* Work supported by the U. S. Atomic Energy Commission, Contract No. AT(11-1)-3077.

CASE 0: $r = n$, $Q > 0$;

CASE 1: $r < n$, $B^T QB > 0$;

CASE k: $r < n$, $B^T A^{T^m} QA^m B = 0$, $m < k-1$, $B^T A^{T^{k-1}} QA^{k-1} B > 0$.

This list is not exhaustive, but serves to illustrate the principal variations of behavior. Case 0 is unusual in practice, so we shall concentrate on CASE 1 and higher cases.

CASE 1

In CASE 1 let us seek an asymptotic expansion in the form

$$\left.\begin{aligned} x &= X(t,\varepsilon) + m(\tau,\varepsilon) + \varepsilon\, n(\sigma,\varepsilon) \\ p &= P(t,\varepsilon) + \varepsilon f(\tau,\varepsilon) + \varepsilon^2 g(\sigma,\varepsilon) \,, \end{aligned}\right\} \tag{5}$$

where each term is to be represented by a series in powers of ε. Here (X,P) is the outer solution, while $(m,\varepsilon f)$ and $(n,\varepsilon^2 g)$ are initial and terminal boundary layer corrections in the stretched time coordinates

$$\tau = \frac{t}{\varepsilon} \,, \quad \sigma = \frac{1-t}{\varepsilon}$$

The boundary layer corrections are required because it is not possible to satisfy all the boundary conditions using the outer outer expansion by itself, as will appear.

Suppose that the outer solution is represented as

$$X(t,\varepsilon) = \sum_{j=0}^{\infty} X_j \varepsilon^j \,, \quad P(t,\varepsilon) = \sum_{j=0}^{\infty} P_j \varepsilon^j \tag{6}$$

Then, equating like powers of ε, we find that

$$B\, R^{-1} B^T P_0 = 0 \,, \quad B\, R^{-1} B^T P_1 = 0 \,.$$

Multiplying on the left by B^T, since R is positive definite, it follows that

$$B^T P_0 = 0 \,, \quad B^T P_1 = 0 \tag{7}$$

Then we find that

$$\dot{X}_j = AX_j + BU_j \tag{8a}$$

$$\dot{P}_j = -QX_j - A^T P_j \tag{8b}$$

where

$$U_j = -R^{-1} B^T P_{j+2} \,, \quad U_{-2} = U_{-1} = 0 \,.$$

Thus for each j, U_j must be such that the constraint

$$B^T P_j = -RU_{j-2} \tag{9}$$

is satisfied. This constraint confines the solution of (8) to a subspace, so that the outer solution cannot satisfy all the boundary conditions of (4). The unsatisfied conditions are eliminated by the boundary layer corrections.

In order to express the constraint (9) in a convenient manner we define the matrix

$$E = I - B(B^TQB)^{-1}B^TQ \tag{10}$$

Then

$$E\,B = 0\;, \quad B^TQ\,E = 0\;, \quad E^2 = E\;. \tag{11}$$

We can now set

$$X_j = EX_j + BU_j\;, \quad P_j = E^TP_j + QBZ_j \tag{12}$$

where

$$V_j = (B^TQB)^{-1}B^TQX_j\;, \quad Z_j = -(B^TQB)^{-1}RU_{j-2}$$

Multiplying (8b) on the left by B^T, it follows from (9) that

$$R\dot{U}_{j-2} = B^TQBV_j + B^TA^TP_j$$

whence

$$V_j = -(B^TQB)^{-1}(B^TA^TE^TP_j + B^TA^TQBZ_j - R\dot{U}_{j-2}) \tag{13}$$

Substituting in (8) we find that

$$\left.\begin{aligned} \frac{d}{dt}(E\,X_j) &= \hat{A}EX_j - \hat{G}\,E^TP_j - H_j \\ \frac{d}{dt}(E^TP_j) &= -\hat{Q}EX_j - \hat{A}^TE^TP_j - K_j \end{aligned}\right\} \tag{14}$$

where

$$\hat{A} = EAE\;, \quad \hat{G} = EAB(B^TQB)^{-1}B^TA^TE^T\;,$$

$$\hat{Q} = E^TQE \geq 0$$

and the forcing terms are

$$H_j = EAB(B^TQB)^{-1}(B^TA^TQBZ_j - R\dot{U}_{j-2})$$

$$K_j = E^TA^TQBZ_j$$

Equations (14) form a Hamiltonian system which can be solved for EX, E^TP with $EX(0)$, $E^TP(1)$ prescribed (Bucy, 1967). Then since V_j is determined by (13), it follows from (12) that X_j satisfies the constraint

$$B^TQX_j = B^TQBV_j \tag{15}$$

It follows from the linearity of the system that the boundary layer corrections must also satisfy (4). Thus

$$\left.\begin{aligned} \frac{dm}{d\tau} &= \varepsilon Am - BR^{-1}B^Tf \\ \frac{df}{d\tau} &= -\,Qm - \varepsilon A^Tf\;, \end{aligned}\right\} \tag{16}$$

where m and f → 0 as τ → ∞. Suppose that m and f are expanded as

$$m(\tau,\varepsilon) = \sum_{j=0}^{\infty} m_j \varepsilon^j \ , \qquad f(\tau,\varepsilon) = \sum_{j=0}^{\infty} f_j \varepsilon^j \tag{17}$$

Then, equating like powers of ε, we find that

$$\frac{dm_j}{d\tau} = - BR^{-1}B^T f_j + A\, m_{j-1} \tag{18a}$$

$$\frac{df_j}{d\tau} = - Q\, m_j \qquad - A^T f_{j-1} \tag{18b}$$

where

$$m_{-1} = f_{-1} = 0$$

Combining (18a) and (18b), we seek decaying solutions of

$$\frac{d^2 m_j}{d\tau^2} = B\, R^{-1} B^T Q m_j + h_{j-1} \tag{19}$$

where

$$h_{j-1} = A\, \frac{dm_{j-1}}{d\tau} + B\, R^{-1} B^T A^T f_{j-1}$$

Let

$$m_j = Em_j + B\, R^{-1/2} M_j \tag{20}$$

where

$$M_j = R^{1/2} (B^T Q B)^{-1} B^T Q m_j \ .$$

Then

$$\frac{d^2}{d\tau^2} Em_j = E\, h_{j-1}(\tau) \tag{21}$$

and

$$\frac{d^2}{d\tau^2} M_j = R^{-1/2} B^T Q\, B\, R^{-1/2} M_j + R^{1/2} (B^T Q B)^{-1} B^T Q h_{j-1} \tag{22}$$

Equation (21) has the decaying particular integral

$$Em_j = \int_{\tau}^{\infty} dr \int_{r}^{\infty} h_{j-1}(s)\, ds \ , \tag{23}$$

Thus $E\, m_j(0)$ is determined. In particular $Em_0(0) = 0$. The homogeneous part of equation (22) has the decaying solution

$$M_j = e^{-C\tau}\, M_j(0) \tag{24}$$

where

$$C = \sqrt{R^{-1/2} B^T Q B R^{-1/2}}$$

from which decaying particular integrals can be constructed by integration for arbitrary $M_j(0)$.

Now to match the initial conditions of (4) we have

$$m_0(0) + X_0(0) = x(0)$$

whence

$$EX_0(0) = Ex(0)$$

providing the initial condition for (14). Also

$B^TQ\,X_0(0)$ is then determined by (15), so $m_0(0)$ must be such that

$$B^TQm_0(0) = B^TQ(x(0) - X_0(0))$$

According to (20) this will be the case if

$$M_0(0) = R^{1/2}(B^TQB)^{-1}B^TQ(x(0)-X_0(0))$$

providing the initial condition for the leading term of the boundary layer. For the higher terms the initial conditions are split in a similar manner. Since $E\,m_j(0)$ is uniquely determined by the requirement for a decaying boundary layer, we have

$$EX_j(0) = -\,Em_j(0)\ , \qquad j > 0\ .$$

Then $B^TQX_j(0)$ is determined, and hence the initial value $M_j(0)$ for the part of the boundary layer correction lying in the range space of B.

The terminal boundary layer correction can be determined in a similar manner. Setting

$$n(\sigma,\varepsilon) = \sum_{j=0}^{\infty} n_j\varepsilon^j\ , \quad g(\sigma,\varepsilon) = \sum_{j=0}^{\infty} g_j\varepsilon^j \tag{25}$$

we find that

$$\left.\begin{aligned} \frac{dn_j}{d\sigma} &= BR^{-1}B^Tg_j - An_{j-1} \\ \frac{dg_j}{d\sigma} &= Qn_j + A^T\,g_{j-1}\ . \end{aligned}\right\} \tag{26}$$

Then

$$\frac{d^2g_j}{d\sigma^2} = Q\ B\ R^{-1}B^Tg_j + k_{j-1} \tag{27}$$

where

$$h_{j-1} = A^T\,\frac{dg_{j-1}}{d\tau} - Q\ A\ n_{j-1}\ .$$

Now we set

$$g_j = E^Tg_j + Q\ B\ R^{-1/2}G_j \tag{28}$$

where

$$G_j = R^{1/2}(B^TQB)^{-1}B^Tg_j\ .$$

Then (27) implies that

$$\frac{d^2}{d\sigma^2}\,E^Tg_j = E^Tk_{j-1} \tag{29}$$

and

$$\frac{d^2G_j}{d\sigma^2} = R^{-1/2}B^TQ\ B\ R^{-1/2}G_j + R^{1/2}(B^TQ\ B)^{-1}B^Tk_{j-1} \tag{30}$$

For a decaying solution $E^Tg_j(0)$ is determined, and hence $E^TP_j(1)$ is determined. Also $B^TP_j(1)$ is determined, providing the terminal value $G_j(0)$.

Thus the complete asymptotic solution can be determined term by term. We observe that the leading term of the outer expansion is just the familiar solution for a singular arc which occurs when $\varepsilon = 0$ (Moylan and Moore, 1971). We also note that near $t = 0$ the dominant term of the control is $(1/\varepsilon)R^{-1}B^Tf_0$, where

$$B\,R^{-1}B^Tf_0 = -\,B\,R^{-1/2}\frac{dM_0}{d\tau} = B\,R^{-1/2}C\,e^{-C\tau}M_0(0)\ .$$

Thus near $t = 0$ the control has the form

$$\frac{1}{\varepsilon}\,R^{-1/2}C\,e^{-C\,t/\varepsilon}\,M_0(0)\ .$$

In the limit as $\varepsilon \to 0$ this reduces to the impulse which is required to make the initial transfer to a singular arc.

HIGHER CASES

In CASE 2 we note that since $Q > 0$ the condition

$$B^TQ\,B = 0$$

implies that

$$Q\,B = 0$$

The appropriate stretching for the initial boundary layer correction proves to be $\tau = t/\mu$, where $\mu = 1/\sqrt{\varepsilon}$. Expanding the outer solution as

$$X(t,\mu) = \sum_{j=0}^{\infty} X_j\mu^j\ , \qquad P(t,\mu) = \sum_{j=0}^{\infty} P_j\mu^j$$

equations (8) hold, where now

$$B^TP_j = R\,U_{j-4}\ , \qquad\qquad U_j = 0\ ,\ j < 0\ . \tag{31}$$

Then (8b) gives

$$B^TA^TP_j = R\,\dot{U}_{j-4} \tag{32}$$

The additional constraint (32) confines the outer solution to a smaller subspace. Consequently a transfer through the space spanned by the columns of B, AB must be accomplished in the initial boundary layer. If the corrections to x and p have the form

$$\frac{1}{\mu}\,m(\tau,\mu) = \frac{1}{\mu}\sum_{j=0}^{\infty} m_j\mu^j\ , \qquad \mu f(\tau,\mu) = \mu\sum_{j=0}^{\infty} f_j\mu^j$$

we find that

$$\frac{d^2m_j}{d\tau^2} = B\,R^{-1}B^TA^Tf_j + A\,\frac{dm_{j-1}}{d\tau}$$

$$\frac{d^2f_j}{d\tau^2} = -\,Q\,A\,m_j - A^T\,\frac{df_{j-1}}{d\tau}$$

The leading term satisfies the homogeneous equation

$$\frac{d^4m_0}{d\tau^4} = -B\,R^{-1}B^TA^TQ\,A\,B\,m_0$$

with the decaying solution

$$m_0 = B\, R^{-1/2} Z i (e^{\rho C\tau} - e^{\bar{\rho} C\tau}) \;, \quad m_0(0) = 0$$

where Z is a constant matrix,

$$\rho = - \frac{1+i}{\sqrt{2}}$$

and

$$C = \sqrt[4]{R^{-1/2} B^T A^T Q\; A\; B\; R^{-1/2}}$$

Then $m_1(\tau)$ lies in the space spanned by the columns of B, AB and can be used to match the boundary conditions at $t = 0$. Near $t = 0$ the dominant term is $(1/\mu)m_0$. Thus x makes an excursion in the range space of B, which approaches the form of a delta function as $\mu \to 0$.

For CASE k we find similarly that the appropriate expansions are in powers of $\mu = \varepsilon^{1/k}$, where the stretching for the the initial boundary layer is $\tau = t/\mu$. Now $B^T A^{T^m} P_j$ is constrained for $m < k$, confining the outer solution to a progressively smaller subspace with increasing k. The initial boundary layer correction has to meet the boundary conditions in the space spanned by the columns of the first k partitions of the controllability matrix

$$[B, AB, \ldots\, A^{n-1}B]$$

and x makes an excursion in the space spanned by the columns of the first k-1 partitions. Thus the increasing thickness of the boundary layer corresponds to the additional work required from the initial control. In the limit the control has a form corresponding to a delta function and its first k-1 derivatives.

REFERENCES

1. A. E. Bryson and Y. C. Ho, Applied Optimal Control, Blaisdell, Waltham, Massachusetts, 1969.

2. R. S. Bucy, Two point boundary value problems of linear Hamiltonian systems, SIAM J. Appl. Math. 15 (1967), pp. 1385-1389.

3. R. E. O'Malley, Jr., and A. Jameson, Singular perturbations and singular arcs, to appear.

4. P. J. Moylan and J. B. Moore, Generalizations of singular optimal control theory, Automatica 7 (1971) pp. 591-598.

A CONCEPTUAL APPROACH TO APPLYING SINGULAR PERTURBATION METHODS TO VARIATIONAL PROBLEMS

ANTHONY J. CALISE
Dynamics Research Corporation
Wilmington, Massachusetts

ROMESH AGGARWAL
Dynamics Research Corp.
Wilmington, Massachusetts

ABSTRACT

A general approach is outlined for using singular perturbation methods on problems in optimal control. Conceptually, the approach can be applied to systems of any order and is particularly useful in obtaining solutions in feedback form.

The approach consists of using time scaling to first reduce the order of the model. The reduced-order model is solved, and the remaining states are then individually modeled by solving separate boundary-layer problems. Since the order of the boundary-layer system is always one, the problem of picking the unknown adjoint values to suppress unstable modes can be avoided by enforcing continuity of the Hamiltonian function in each layer. The general approach has both conceptual and practical value in understanding the dependence of the optimal solution on initial conditions, path and control variable constraints and terminal constraints. An aircraft control problem is treated to illustrate the usefulness of the approach. Three-dimensional numerical results are given along with comparisons to the optimal solutions.

INTRODUCTION

Numerous papers have appeared in the literature which treat the use of singular perturbation methods in initial-value and two-point boundary-value problems. In general the techniques used are based on boundary-layer solution methods of Vasil'eva [1] and others. The key advantage in applying singular perturbations lies in the resulting reduction of order, so that the solution of a high-order system can be reduced to the solution of several systems of lower order.

Recently, Calise [2] has discussed additional advantages that result from the use of singular perturbation methods on problems in optimal control. In particular, these methods are useful in solving state-constrained variational problems and in obtaining solutions in feedback form. In practice, solutions in this form are of paramount importance in both real-time and non-real time applications [3].

The greatest difficulty encountered in applying singular perturbation methods lies in finding matching boundary-layer solutions which are needed to correct for discrepancies in the so called "fast variables" in the free-stream solution. These discrepancies generally occur at the initial and terminal times. In the initial boundary layer system, the initial values of the adjoint variables are free and must be picked to both suppress unstable modes and to insure that the boundary-layer corrections asymptotically approach zero. Kelley [4] has shown for a simple Mayer variational problem, that the strengthened Legendre-Clebsch condition at the end-points is essential for success. Sannuti and Reddy [5] have illustrated a solution for several specialized cases. However, for the general nonlinear case, finding matched asymptotic solutions is a difficult problem.

The intent of this paper is to outline an approach which avoids the difficulty of picking the unknown parameters in the boundary-layer system. This is done on the basis of multiple time-scale separation of the state equations [6], and enforcing the requirement that the Hamiltonian be continuous across the separate boundary layers. Furthermore, it is shown that whenever possible, the zero-order free-stream model should be reduced to the point where a feedback solution is possible, or where the open-loop free-stream solution can be computed readily. This allows the total solution to be updated based on a feedback of the true current state, thus improving the accuracy of the zero-order solution and reducing the need for higher order expansion terms as in [5].

OUTLINE OF THE APPROACH

Consider the system of equations

$$\dot{x} = f(x, y, u, t) \qquad (1)$$

$$\epsilon_i \dot{y}_i = g_i(x, y, u, t) \quad i = 1, 2 \cdots m \qquad (2)$$

where x, y and u are n, m and p dimensional vectors respectively. The parameters ϵ_i are introduced as time-scaling parameters, so that the individual elements of y are ordered on separate time scales in accordance with the relative speeds with which these variables can change their magnitudes. The elements ϵ_i are taken such that

$$\lim_{\epsilon_i, \epsilon_{i+1} \to 0} \left({}^{\epsilon_{i+1}}/_{\epsilon_i} \right) = 0 \qquad (3)$$

It is desired to minimize a terminal function of the states[†]

$$J = \phi(x, y) \qquad (4)$$

subject to an arbitrary set of state and/or control variable constraints

$$\psi_1(x, y, u, t) \le 0 \qquad (5)$$

and an arbitrary set of terminal constraints

$$\psi_2[x(t_f), y(t_f), t_f] = 0 \qquad (6)$$

Free-Stream Solution

Taking the limit as the ϵ_i go to zero, Equations (1) and (2) become

$$\dot{x} = F(x, y_1, u, t) \qquad (7)$$

$$0 = G_1(x, y_1, u, t) \qquad (8)$$

where the elements y_i for i = 2, 3 $\cdots$ m are here assumed to be eliminated as functions of x, u, y_1 and t using†

[†] An integral function of the states and control can, in general, be included as part of J in the development that follows.

$$g_i = 0 \qquad i = 2, 3 \cdots m \tag{9}$$

The necessary conditions for optimality can now be written as

$$\dot{x} = F\,(x,\ y_1,\ u,\ t), \quad x(o) = x_o \tag{10}$$

$$\dot{\lambda}_x = -\frac{\partial H_o}{\partial x}, \quad \lambda_x(t_f) = \frac{\partial \phi}{\partial x} + \nu_x^T \frac{\partial \psi_2}{\partial x} \tag{11}$$

$$\frac{\partial H_o}{\partial y_1} = 0 \qquad \frac{\partial H_o}{\partial u} = 0 \tag{12}$$

where

$$H_o = \lambda_x F + \bar{\lambda}_{y_1} G_1 + \nu^T \Psi(x,\ y_1,\ u,\ t) \tag{13}$$

and $\bar{\lambda}_{y_1}$ is chosen to enforce the constraint in Equation (8). It should be noted that the non-linear relationship between x, y_1 and u in Equation (8) can be used to avoid the difficulties normally encountered whenever u appears linearly (singular variational problems) or when some of the element of x and y are constrained independent of u (state constrained variational problems).

The solution of Equations (10) through (13) will yield a zero-order free-stream solution consisting of $\bar{x}$, $\bar{\lambda}_x$, $\bar{\lambda}_{y_1}$, $\bar{\nu}$, $\bar{u}$ and $\bar{y}_1$. As mentioned in the introduction, from the standpoint of application and in order to minimize the need for higher order correction terms, every attempt should be made to sufficiently reduce the free-stream solution so that either a feedback solution results or the open-loop solution can readily be computed. An example of the later case is given in this paper. For examples of non-linear problems where the free-stream solution can be reduced directly to a feedback form see Reference [2]. As will be shown, all the boundary-layer correction terms are in feedback form.

Boundary Layer Solution

Following the usual procedure, the time transformations

$$\tau_i = \frac{t}{\epsilon_i} \tag{14}$$

and

$$\sigma_i = \frac{t_f - t}{\epsilon_i} \tag{15}$$

are used to derive the zero-order boundary-layer correction terms. In this section, the i=1 boundary layer solution is derived for the initial time and for a scaler u.[†] The derivation for the remaining "outer boundary layers" follows by induction by considering the zero-order free-stream plus i=1 boundary-layer solution derived here to act as a new free-stream solution for the i=2 boundary layer. In the ith boundary layer analysis, the state variable from the i+1 layer acts as a control variable as was the case with y_1 in Equation (12). The dynamics for attaining the resulting optimal state value ($\bar{y}_{i+1}$) are then investigated in the i+1 layer until the outer most layer is reached. Then only the original u vector appears as a control.

---

† We assume here that this function exists and is unique.

Using the transformation in Equation (14) and letting ϵ_1 go to zero, the necessary conditions for optimality become

$$x' = 0 \tag{16}$$

$$y' = G_1(\bar{x}, y_1, u, t), \quad \begin{cases} y(o) = y_o \\ y(\infty) = \bar{y}_1 \end{cases} \tag{17}$$

$$\lambda'_{y_1} = -\frac{\partial H_1}{\partial y_1}, \quad \lambda_{y_1}(\infty) = \bar{\lambda}_{y_1} \tag{18}$$

$$\frac{\partial H_1}{\partial u} = 0 \tag{19}$$

where

$$H_1 = \bar{\lambda}_x F(\bar{x}, y_1, u, t) + \lambda_{y_1} G_1 + \nu^T \Psi(\bar{x}, y_1, u, t) \tag{20}$$

Since our main interest lies in driving y_1 to $\bar{y}_1$, it is convenient to think of G_1 as a control variable replacing u so that H_1 can be represented as

$$H_1 = \bar{\lambda}_x F_1(\bar{x}, y_1, G_1, t) + \lambda_{y_1} G_1 + \nu^T \Psi_1(\bar{x}, y_1, G_1, t) \tag{21}$$

Expanding H_1 in a Taylor series about $\bar{y}_1$ and $G_1=0$, we have

$$H_1 = H_o + \frac{\partial^2 H_o}{\partial y_2^2}\Bigg|_{\substack{y_1=\bar{y}_1 \\ G_1=0}} \frac{\delta y_1^2}{2} + \bar{\lambda}_x \Bigg[\frac{\partial F_1}{\partial G_1}\Bigg|_{G_1=0} G_1$$

$$+ \frac{\partial^2 F_1}{\partial G_1^2}\Bigg|_{G_1=0} \frac{G_1^2}{2}\Bigg] + \lambda_{y_1} G_1 + \nu^T \Bigg[\frac{\partial \Psi_1}{\partial G_1}\Bigg|_{G_1=0} G_1 + \frac{\partial^2 \Psi_1}{\partial G_1^2}\Bigg|_{G_1=0} \frac{G_1^2}{2}\Bigg] = H_o(o) \tag{22}$$

where

$$\delta y_1 = y_1 - \bar{y}_1 \tag{23}$$

Note that the first variation of H_o with respect to y_1 is zero from Equation (12) and that the variations of F_1 and Ψ_1 are evaluated for the current value of y_1. Taking the variation of H_1 with respect to G_1 yields

---

[†] The results here can be extended to the case where u is a vector.

$$\bar{\lambda}_x\left[\frac{\partial F_1}{\partial G_1} + \frac{\partial^2 F_1}{\partial G_1^2} G_1\right] + \lambda_{y_1} + \nu^T\left[\frac{\partial \Psi}{\partial G_1} + \frac{\partial^2 \Psi}{\partial G_1^2} G_1\right] = 0 \qquad (24)$$

Solving for λ_{y_1} above and substituting back into Equation (22) results in the following expression G_1

$$G_1 = -\sqrt{\frac{\partial^2 H_o}{\partial y_1^2} \Big/ \left[\nu^T \frac{\partial^2 \Psi}{\partial G_1^2} + \bar{\lambda}_x \frac{\partial^2 F_1}{\partial G_1^2}\right]}\ \delta y_1 \qquad (25)$$

where ν is chosen to ensure that the constraints in Equation (5) are met.

Equation (25) shows that the rate of y_1 should be proportional of $y_1 - \bar{y}_1$ or that y_1 approaches $\bar{y}_1$ exponentially in time. As y_1 approaches $\bar{y}_1$ and G_1 approaches zero, it is apparent from Equations (13) and (22) that λ_{y_1} and ν will approach $\bar{\lambda}_{y_1}$ and $\bar{\nu}$. Thus, the solution satisifies the matching requirement. The positive root in Equation (25) is taken in the terminal boundary layer where y_1 will depart eponentially from $\bar{y}_1$ to meet the terminal constraint set.

OPTIMAL AIRCRAFT CONTROL FOR DELIVERING AN AIR-TO-AIR MISSILE

The equations of motion defining the attacking aircraft dynamics are [See Fig. 1]:

$$\dot{x}_r = V\cos\gamma\ \sin(\phi_T - \phi) + \dot{\phi}_T y_r \qquad (26)$$

$$\dot{y}_r = V\cos\gamma\ \cos(\phi_T - \phi) - \dot{\phi}_T x_r - V_T \cos\gamma_T \qquad (27)$$

$$\dot{E} = (TH - D)\, V/W \qquad (28)$$

$$\epsilon_1 \dot{\phi} = gL\sin\mu / WV\cos\gamma \qquad (29)$$

$$\epsilon_2 \dot{h} = V\sin\gamma \qquad (30)$$

$$\epsilon_3 \dot{\gamma} = g[L\cos\mu/W - \cos\gamma]/V \qquad (31)$$

where E, h, γ and W are the total energy, altitude, flight path angle and weight of the attacking aircraft. The target motion is defined a-priori by a velocity (V_T) and turn-rate history ($\dot{\phi}_T$). The objective of the attacking aircraft is to maximize missile kill probability PK which is given as a function of relative position (x_r, y_r. h_r), at the time of missile launch t_f. The attacker's controls are lift (L) and the bank angle (μ).

The following aircraft constraints are imposed:

$$L/W \leq G_{max} \tag{32}$$

$$L \leq 1/2\, \rho V^2 S_{ref}\, C_{\ell\, max} \tag{33}$$

$$h \geq h_{min} \tag{34}$$

$$V = [2g(E-h)]^{1/2} \leq V_{Limit}(h)^{\dagger} \tag{35}$$

The terminal constraint for the attacker is:

$$\beta(t_f) = 0 \tag{36}$$

This constraint is imposed to allow the missile seeker to lock-on at a zero boresight condition immediately prior to launch.

Two Dimensional Singular Perturbation Solution

The following solution is obtained for two-dimensional target and attacking aircraft maneuvers. This is equivalent to obtaining a free-stream and first-boundary layer solution for the three-dimensional problem. For $\epsilon_2 = \epsilon_3 = 0$, Equations (30) and (31) give

$$\gamma = 0 \tag{37}$$

$$L/W = \sec \mu \tag{38}$$

and equation (29) becomes

$$\epsilon_1 \dot{\phi} = g \tan \mu / V \tag{39}$$

Free Stream Solution

Taking the limit as ϵ_1 goes to zero and using the necessary conditions for optimality given by equations (10)-(13) we obtain

$$H_o = \lambda_x [V \sin(\phi_T - \phi) + \dot{\phi}_T y_r] + \lambda_y [V \cos(\phi_T - \phi) - \dot{\phi}_T x_r - V_T] + \lambda_E (TH-D) \frac{V}{W} \tag{40}$$

$$\dot{\lambda}_x = -\frac{\partial H_o}{\partial x_r} = \dot{\phi}_T \lambda_y \tag{41}$$

$$\dot{\lambda}_y = -\frac{\partial H_o}{\partial y_r} = -\dot{\phi}_T \lambda_x \tag{42}$$

$$\dot{\lambda}_E = -\frac{\partial H_o}{\partial E} \tag{43}$$

$$\frac{\partial H}{\partial \phi} = 0 \tag{44}$$

Equation (44) yields

$$\tan(\phi_T - \phi) = -\frac{\lambda_x}{\lambda_y} \tag{45}$$

---

† The velocity limit here is a composite of maximum dynamic pressure and maximum mach.

Equations (41), (42) and (45) imply that the free-stream solution consists of a dash. This can be seen by differentiating both sides of Equation (45) and making the appropriate substitutions from Equations (41) and (42).

At the terminal time t_f

$$\lambda_x(t_f) = \frac{\partial PK}{\partial x_r(t_f)}, \quad \lambda_y(t_f) = \frac{\partial PK}{\partial y_r(t_f)} \tag{46}$$

$$\lambda_E(t_f) = 0 \tag{47}$$

Equations (45) and (46) define the stopping condition for the optimal free stream solution.

Now let us assume

1. A parabolic polar drag

$$D = 1/2\,\rho V^2 S_{ref} C_{Do} + \frac{KL^2}{1/2\rho V^2 S_{ref}} \tag{48}$$

2. Contours of constant kill probability are circles with linear variation with respect to distance from the center.

3. Target velocity V_T, turn rate $\dot{\phi}_T$ and the attacking aircraft energy remain constant in the free stream.[†]

Under these assumptions the free stream is obtained by geometric considerations for both the fixed-time and free-time problems. The target and thus the center of the PK contours move in a circle. For a fixed-time problem the free-stream solution is a dash towards the terminal position of the center of PK contours. For the free-time problem, the optimal final t_f must satisfy the condition

$$[\,(x_o(t_f) - x(o))^2 + (y_o(t_f) - y\,(o))^2\,]^{1/2} = V * t_f \tag{49}$$

where (x_o, y_o) is the known position of PK contour center as a function of time in the fixed frame and (x, y) is the attacker's position in the fixed frame. The optimal $\bar{\phi}$ is given by

$$\bar{\phi} = \tan^{-1}\left[\frac{y_o(t_f)-y\,(0)}{x_o(t_f)-x\,(0)}\right] \tag{50}$$

When the attacker's velocity is not sufficient to reach the center of the PK region, t_f and $\bar{\phi}$ are computed to yield the point of closest approach.

Under assumption 2,

[†] This assumption is consistent with the fact that the pilot cannot predict changes in V_T and $\dot{\phi}_T$.

$$\lambda_x(t_f) = \frac{\sin(\phi_T(t_f)-\bar{\phi})}{R}, \qquad \lambda_y(t_f) = \frac{\cos(\phi_T(t_f)-\bar{\phi})}{R} \tag{51}$$

where R is the outer radius of PK region. Also, under assumption 1

$$\dot{\lambda}_E = -\frac{g}{VR} - \left[TH - 1.5\,\rho V^2 S_{ref}\, C_{Do} + \frac{2KW^2}{\rho V^2 S_{ref}}\right] \frac{g}{V_o W} \lambda\, E \tag{52}$$

Using Equations (41), (42), (47), (51) and (52) the adjoint variables are integrated analytically to obtain their values at the initial time.

First Boundary-Layer Solution

Incorporating the time transformation $\tau = t/\epsilon_1$ results in the boundary-layer equations for the heading transition.

Using the necessary conditions for the boundary-layer solution given by equations (16) - (20),

$$H_1 = \bar{\lambda}_{x_{ro}} [V_o \sin(\phi_T - \phi) + \dot{\phi}_T\, y_{ro}] + \bar{\lambda}_{y_{ro}} [V_o \cos(\phi_T - \phi) - \dot{\phi}_T\, x_{ro} - V_T] + \bar{\lambda}_{Eo} \left[TH - 1/2\, \rho V^2 S_{ref} C_{Do} - \frac{KW^2}{1/2\, \rho V_o^2 S_{ref}} \left(1 + \left(\frac{V_o}{g} \dot{\phi}\right)^2\right)\right] \frac{V_o}{W} + \lambda_\phi \dot{\phi} = H_o(o) \tag{53}$$

Note that $\dot{\phi}$ has been substituted as a control variable as suggested by Equation (21). Defining $\delta\phi = \phi - \bar{\phi}$ and expanding Equation (53) about $\bar{\phi}$, we have

$$H_1 = H_o + V_o[\lambda_{xo} \sin(\phi_T - \bar{\phi}) + \lambda_{yo} \cos(\phi_T - \bar{\phi})] \frac{\delta\phi^2}{2} - \bar{\lambda}_{Eo} \frac{KW}{1/2\, \rho V_o S_{ref}} \left(\frac{V_o \dot{\phi}}{g}\right)^2 + \lambda_\phi\, \phi = H_o(o) \tag{54}$$

taking the variation of H_1 with respect to $\dot{\phi}$ yields

$$\dot{\phi} = -\left[\frac{\rho g^2 S_{ref}}{4KWR\, \bar{\lambda}_{Eo}}\right]^{1/2} \delta\phi \tag{56}$$

The original control variables are then given by

$$\mu = \tan^{-1}\left[\frac{V\dot{\phi}}{g}\right] \tag{57}$$

and

$$L = W \sec\mu \tag{58}$$

Equation (56) implies that the attacking aircraft heading approaches the free stream heading ϕ exponentially. The positive root in Equation (56) holds for the heading transition in the terminal boundary layer to meet the terminal constraint in Equation (36).

Attacking aircraft controls obtained from Equations (57) and (58) are used to drive the actual system for a small time interval. The new states of the system are then used to update the free stream solution thus obtaining the total solution in the feedback form.

Discussion of Results

Figure 2 illustrates a 2-dimensional solution where the free-stream $\bar{\phi}$ and time are not updated. The solution trajectory does not meet the terminal heading constraint because the free-stream solution time is too small for the total heading transition required. The kill probability is zero. Figs. 3 and 4 illustrate the solution trajectory and turn-rate history for the same problem where the free-stream solution is updated. The dashed line trajectory is the optimal trajectory for this problem. The updated solution is in good agreement with the optimal solution. Figs. 5 and 6 illustrate a 3-dimensional updated-solution trajectory vs. an optimal 3-dimensional solution trajectory for a free time problem. Again, these results are in good agreement with optimal solution.

SUMMARY

A general approach has been outlined for applying singular perturbation methods to variational problems. The approach avoids the matching problem of picking the unknown parameters in the boundary-layer analysis. Updating the free-stream solution greatly improves the accuracy of the method and allows the free-time problem to be solved.

REFERENCES

1. A. B. Vasil'eva, "Asymptotic Behavior of Solutions to Certain Problems Involving Non-Linear Differential Equations Containing a Small Parameter Multiplying the Highest Derivatives", Russian Math Surveys, Vol. 18, No. 3, pp. 13-81, 1963.
2. A. J. Calise, "On the Use of Singular Perturbation Methods in Variational Problems", JACC, Columbus, Ohio, June 20-22, 1973.
3. A. J. Czuchry, A. J. Calise, "Energy Management Rules For Turning Flight", AIAA 11th Aerospace Sciences Meeting, Washington, D. C., January 10-12, 1973.
4. H. J. Kelley, "Singular Perturbations for a Mayer Variational Problem", AIAA Journal, Vol. 8, No. 6, pp. 1177-1178, 1970.
5. P. Sannuti, P. B. Reddy, "Optimization of High-Order Time Delay Systems by the Method of Asymptotic Expansions", Proc. Int. Conf. on Cyber, and Soc., pp 253-258, 1972.
6. H. J. Kelley, "Flight Path Optimization with Multiple Time Scales", J. Aircraft, Vol. 8, No. 4, pp. 238-240, 1971.

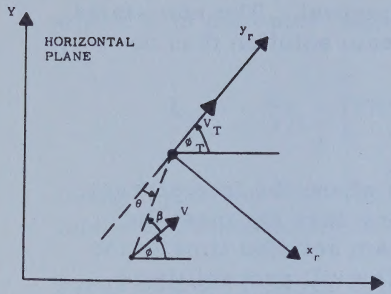

Y
HORIZONTAL PLANE
y_r
V_T
ϕ_T
θ
β
V
ϕ
x_r

FIG. 1 3-D ATTACKER DYNAMICS

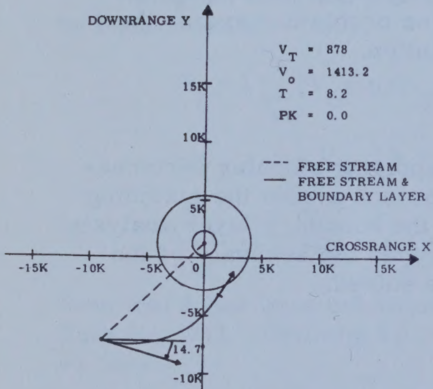


FIG. 2 2-D SOLUTION WITHOUT FREE STREAM UPDATE

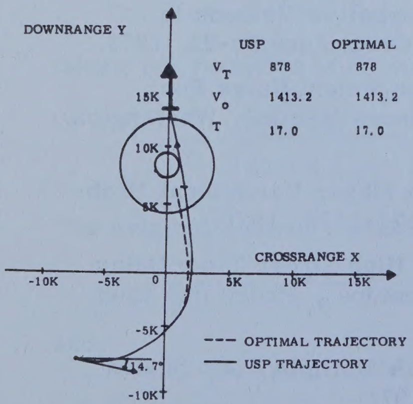


FIG. 3 COMPARISON OF OPTIMUM TRAJECTORIES VS. USP STRATEGIES

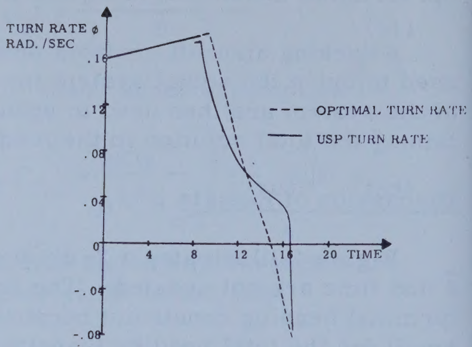


FIG. 4 COMPARISON OF OPTIMUM TURN RATE HISTORY VS. USP TURN RATE HISTORY

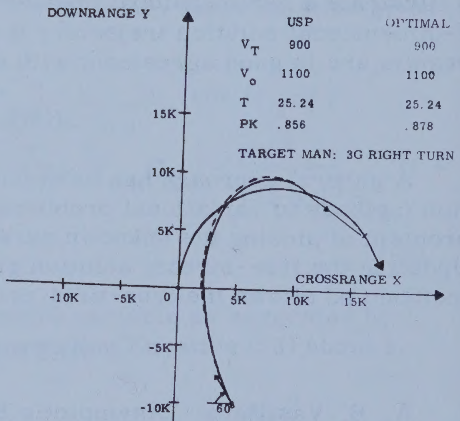


FIG. 5 COMPARISON OF 3-D OPTIMAL TRAJECTORY VS. 3-D-UPS TRAJECTORY

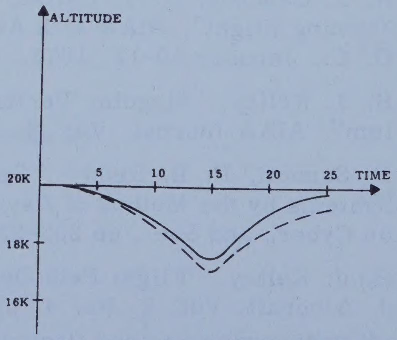


FIG. 6 COMPARISON OF OPTIMAL ALTITUDE HISTORY VS. UPS ALTITUDE HISTORY

ASYMPTOTIC STABILITY IN NETWORKS WITH PARASITICS

BY FRANK HOPPENSTEADT*
Courant Institute of Mathematical Sciences, New York University,
New York 10012

Abstract

An important question is how the presence of parasitics can influence the large-time stability of an electrical network. Many investigations of this problem have been based on the construction of Liapunov functions, a method which can become quite tedious. However, frequently much more information can be obtained under the same conditions by using a well known perturbation method to construct approximate solutions of the problem. These lead to useful approximations to the domain of stability, to the large-time state of the system, and in asymptotically steady problems, to the system's steady state.

Systems of differential equations with a small parameter appearing in the derivative terms, such as

$$dx/dt = f(t,x,y,\varepsilon), \qquad \varepsilon\, dy/dt = g(t,x,y,\varepsilon) \tag{P_ε}$$

where ε is a small positive parameter and x and y are m- and n- vectors, respectively, arise frequently in a wide variety of applications. In particular, in electrical networks, the parameter ε may account for the presence of various "parasitics" in the network.

An important problem, and one which several investigators have studied [1-5], is concerned with how the presence of parasitics influences the large-time stability of the network. Most previous analyses of this problem have involved the construction of Liapunov functions to establish the network's stability. From a practical point of view, that method gives only a "yes" or "maybe not" answer to the question "Is the network stable"? In order to obtain more precise information in a given situation, such as sharp approximations to the asymptotic limit of the system, more must be known about the Liapunov functions used. In some cases, such as [2], this is fairly straightforward, in other cases, as in [1], it is quite tedious.

It is possible that through the addition of some restrictions to the general case considered in [1], which still allow for the treatment of a

---

*This Research was supported by the U. S. Army Research Office-Durham, under Grant No. DA-ARO-D-31-124-72-G47.

very broad class of problems, a great deal more information can be obtained about the large time behavior of the system. This may be done by taking advantage of the method of matched asymptotic expansions, which is widely used to analyze problems like (P_ε). By means of this method, it is possible to determine not only whether the system is stable, but in addition, to compute an approximation to the large time state, and in the asymptotically steady problem, to the steady state.

The details of this application are given in [6], and only the asymptotically steady case will be briefly described here. The several assumptions to be made about the system (P_ε) are denoted by I, II, III.

(I) The reduced problem ($\varepsilon = 0$ in (P_ε)),

$$dx/dt = f(t,x,y,0)\ , \quad 0 = g(t,x,y,0) \qquad (P_0)$$

has a smooth solution $x = x_0(t)$, $y = y_0(t)$, existing for $0 \leq t < \infty$. Moreover, the functions f and g are assumed to be smooth functions, say having derivatives with respect to t,ε and the components of x and y up to order $N+2$, in some uniform tube about $x = x_0(t)$, $y = y_0(t)$ for ε near zero. Finally, this solution has a finite limit $(x_0(\infty), y_0(\infty))$ as $t \to \infty$, and the functions f,g and their derivatives, when evaluated at $x = x_0(t)$, $y = y_0(t)$, $\varepsilon = 0$, have finite limits as $t \to \infty$.

(II) The eigenvalues of the Jacobian matrix

$$g_y(t,\ x_0(t),\ y_0(t),\ 0)$$

all have real parts less than some fixed negative number for all $0 \leq t < \infty$. Here g_y denotes the matrix having components $(\partial g_i/\partial y_i)$.

(III) The matrix

$$A(t) = (f_x - f_y g_y^{-1} g_x)\ (t, x_0(t),\ y_0(t), 0)$$

approaches a stable matrix $A(\infty)$ as $t \to \infty$; i.e., $A(\infty)$ is an $m \times m$ - matrix whose eigenvalues satisfy $\mathrm{Re}\ \lambda < 0$.

These assumptions have several implications. First, the variable y can be eliminated from the system (P_0). In fact, there is a smooth function, $\phi(t,x)$, which satisfies

$$g(t,x,\phi(t,x)\ ,\ 0) = 0$$

for x near $x_o(t)$ and $0 \leq t < \infty$. Moreover, $y_o(t) = \phi(t,x_o(t))$ for $t \geq 0$.

Next, the solution $x = x_o(t)$ of

$$dx/dt = f(t,x,\phi(t,x),0) \tag{P_o'}$$

is asymptotically stable. In fact, the linear part of this problem involves exactly the matrix $A(t)$, and this approaches a stable matrix, as $t \to \infty$. Therefore, there is a domain of stability, say D, for this solution such that if $\xi^* \in D$, then the solution of (P_o') which satisfies $x(0) = \xi^*$, exists for all $t \geq 0$ and approaches $x_o(t)$ as $t \to \infty$.

Finally, the solution $Y = 0$ of the "fast time" problem, obtained from (P_ε) by setting $\tau = t/\varepsilon$ and $\varepsilon = 0$,

$$dY/d\tau = g(0,\xi,\phi(0,\xi) + Y,\ 0) \tag{F_o}$$

is asymptotically stable for each choice of ξ near $x_o(0)$ because of condition (II). Thus, for each such ξ there is a domain of attraction, E_ξ, which we write in such a way that for

$$Y(0) + \phi(0,\xi) \in E_\xi\ ,$$

the corresponding solution $Y(\tau)$ of (P_o) exists for $\tau \geq 0$ and approaches zero as $\tau \to \infty$.

The following result is proved in [6].

Theorem: Let conditions I, II and III be satisfied. If $(\xi(\varepsilon),\eta(\varepsilon))$ are smooth functions of ε with $\xi(0) \in D$ and $\eta(0) \in E_{\xi(o)}$, then for small $\varepsilon > 0$, the problem (P_ε) has a unique solution $(x(t,\varepsilon),y(t,\varepsilon))$ for $0 \leq t < \infty$ which satisfies

$$x(0,\varepsilon) = \xi(\varepsilon), \qquad y(0,\varepsilon) = \eta(e)\ .$$

Moreover,

$$\lim_{t\to\infty} (x(t,\varepsilon),\ y(t,\varepsilon)) = (x^*(\varepsilon),y^*(\varepsilon))$$

where $(x^*(\varepsilon),y^*(\varepsilon))$ are smooth functions of ε at $\varepsilon = 0$

and satisfy

$$0 = f(\infty, x^*(\varepsilon), y^*(\varepsilon),\varepsilon) \ , \quad 0 = g(\infty, x^*(\varepsilon), y^*(\varepsilon),\varepsilon) \ ,$$

$$x^*(0) = x_o(\infty) \ , \quad y^*(0) = y_o(\infty) \quad .$$

It follows from this result that the limit $(x^*(\varepsilon),y^*(\varepsilon))$ can be approximated by applying the implicit function theorem. In fact, we would have

$$(x^*(\varepsilon),y^*(\varepsilon)) = (x_o(\infty),y_o(\infty)) + \sum_{r=1}^{N} (x_r(\infty),y_r(\infty))\varepsilon^r + 0(\varepsilon^{N+1})$$

where

$$x_r(\infty) = - A^{-1}(\infty) \ R_r(\infty)$$

and $$y_r(\infty) = - g_y^{-1}(\infty) \ [g_x(\infty) \ x_r(\infty) + g_r(\infty)] \quad .$$

Here $g_y(\infty)$, etc., denotes $\lim_{t\to\infty} g_y \ (t,x_o(t),y_o(t),0)$, etc., and the vectors $R_r(\infty)$ and $g_r(\infty)$ can be determined in a straightforward way.

More importantly, this result shows that any initial data lying in the set $\{(\xi,E_\xi) : \xi \in D\}$ approaches this value asymptotically as $t \to \infty$. Thus, the result gives an estimate of the domain of attraction. The proof of this result rests on the development of the method of matched asymptotic expansions given in [7].

This approach may be generalized directly to study the stability of systems involving several parameters, which are amenable to the method of matched asymptotic expansions. Such a class of problems is described in [7]. In these cases the large time behavior of solutions is again given by an expansion for the outer solution of the problem which can be obtained directly from the system.

BIBLIOGRAPHY

1. F. Hoppensteadt, (1968), "Asymptotic stability in singular perturbation problems", J. Diff. Eqns. (4), 350-358.
2. R. R. Wilde and P. V. Kokotovic, "Stability of singularly perturbed systems and networks with parasitics", IEEE Auto Control (April,1972).
3. C. A. Desoer and M. J. Shensa, "Networks with very small and very large parasitics: natural frequencies and stability", Proc. IEEE, vol. 58, pp. 1933-1938 (1970).
4. M. J. Shensa, "Parasitics and the stability of equilibrium points of nonlinear networks", IEEE Trans. Circuit Theory, vol. CT-18, pp. 481-484 (1971).
5. C. A. Desoer, "Singular perturbation and bounded-input bounded-state stability", Elec. Lett., vol. 6, pp. 496-497 (1970).
6. F. Hoppensteadt, "Asymptotic stability in singular perturbation problems II", J. Diff. Eqns., (in press).
7. F. Hoppensteadt (1971), "Properties of solutions of ordinary differential equations with small parameters", Comm. Pure Appl. Math. (XXIV), 807-840.

OPTIMIZATION OF A COUPLED-CORE NUCLEAR REACTOR SYSTEM BY THE METHOD OF ASYMPTOTIC EXPANSIONS

P. B. Reddy*
Dynamics Research Corporation
Wilmington, Massachusetts

P. Sannuti
Rutgers University
New Brunswick, New Jersey

ABSTRACT

A two-core coupled nuclear reactor system described by an eighth order nonlinear time delay model is reduced to a second order ordinary model by perturbing a small scalar parameter. An asymptotic series solution is constructed with respect to the scalar parameter for the eighth order time delay model by using the solution of the reduced second order ordinary model. Computationally, the asymptotic approximation method is compared with the second variation method and is shown to be superior.

INTRODUCTION

Many practical engineering systems require a high number of dynamic equations to represent their behavior. Pure time delays are commonly encountered in many such systems. It is well known that the present day optimal control synthesis procedures require a prohibitive amount of computation for such complex systems. In an earlier paper [1], an optimal asymptotic series solution is developed for a class of high-order nonlinear time delay models by using solutions of lower order ordinary models. The present paper illustrates the applicability and computational simplicity of the asymptotic method on a coupled-core nuclear reactor system.

In the years ahead nuclear power is expected to be one of the main sources of energy for an ever-increasing energy demand. The energy that could be generated from a pound of nuclear fuel is enormous compared to that generated by a pound of fossil fuel. For certain applications, such as the use of nuclear reactors to supply the propulsive power for submarines, naval ships, military aircraft and space rockets, the tremendous saving in the weight of the fuel will more than compensate for the mass of the reactor and its accessory equipment. Under these circumstances the cost of power is of secondary importance and its availability is of prime importance. The increasing use of nuclear power sources for many applications has brought about an increasing interest in optimal nuclear reactor control. The power output of a nuclear reactor system can be increased by coupling a number of individual reactor cores together. Also, in propulsion applications, the concept of clustering a number of individual nuclear propulsion engines has been gaining serious consideration. This is because it offers versatility in the application of a single standard engine for a variety of mission requirements, and also an economy in the nuclear engine development effort. When two or more cores are coupled together, there is a mutual exchange of neutrons among the cores. Thus the neutron behavior of a specific core depends upon the past history of all the other cores. Also the number of system equations is increased with each additional coupled reactor. In general, to decrease the number of system equations, either the effect of delayed neutrons or the effect of temperature variations are neglected [2]. Here we take into account all these effects and obtain an eighth order time delay system of equations for a two core coupled reactor. This eighth order time delay system is reduced to a second order ordinary system and then the method of asymptotic expansions [1, 3, 4, 5] is utilized to find an approximate optimal control.

---

*Formerly with the Department of Electrical Engineering, Rutgers University, New Brunswick, New Jersey.

COUPLED-CORE REACTOR EQUATIONS

A two-core coupled reactor system is shown in Fig. 1. As mentioned earlier, because of the neutron leakage from each core, the behavior of a specific core depends upon the past history of the other core. Thus the coupled system equations contain variables whose arguments are delayed in time. The following assumptions are made in the development of the system equations: (a) the coupling effect in each core appears in the form of a source from the other core, (b) the external reactivity ρ_e in cores 1 and 2 is assumed to be controlled by servomotors as shown in Fig. 2. Now the two core coupled system equations are [2, 6],

$$\begin{aligned}\dot{n}_1 &= -\frac{\alpha_{21}}{\ell} n_1(t) + \frac{\alpha_{12}}{\ell} n_2(t-\tau_{12}) - \frac{\beta}{\ell} n_1(t) + \nu c_1(t) \\ &\quad - \frac{\alpha_T}{\ell} T_1(t) n_1(t) + \frac{\rho_{e_1}}{\ell} n_1(t) \qquad (1a)\end{aligned}$$

$$\begin{aligned}\dot{n}_2 &= -\frac{\alpha_{12}}{\ell} n_2(t) + \frac{\alpha_{21}}{\ell} n_1(t-\tau_{21}) - \frac{\beta}{\ell} n_2(t) + \nu c_2(t) \\ &\quad - \frac{\alpha_T}{\ell} T_2(t) n_2(t) + \frac{\rho_{e_2}}{\ell} n_2(t) \qquad (1b)\end{aligned}$$

$$\dot{c}_1 = \frac{\beta}{\ell} n_1(t) - \nu c_1(t) \tag{1c}$$

$$\dot{c}_2 = \frac{\beta}{\ell} n_2(t) - \nu c_2(t) \tag{1d}$$

$$\dot{T}_1 = \frac{1}{(MC_r)_1} n_1(t) - a_1 T_1(t) \tag{1e}$$

$$\dot{T}_2 = \frac{1}{(MC_r)_2} n_2(t) - a_2 T_2(t) \tag{1f}$$

$$\dot{\rho}_{e_1} = k_1 \rho_{e_1}(t) + k_2 u_1(t) \tag{1g}$$

$$\dot{\rho}_{e_2} = k_3 \rho_{e_2}(t) + k_4 u_2(t) \tag{1h}$$

where

n_i = power level in core i
c_i = delayed neutron power level in core i
T_i = temperature in core i
ρ_{ei} = external control reactivity in core i
α_{ij} = coupling coefficient of proportionality, $i \neq j$
α_T = reactivity temperature coefficient
τ_{ij} = delay time associated due to neutron transport, $i \neq j$
ℓ = mean prompt neutron generation time
β = fraction of the total neutrons that are delayed in time
ν = inverse of the mean life time of the delayed neutrons
$(MC_r)_i$ = product of the mass and specific heat of the i'th core
a_i = inverse of the heat exchanger time constant
k_2, k_4 : gain constants of the controllers
k_1, k_3 : inverse of the controller's time constants
u_i = control input to the servomotor i

and the dot, •, indicates the derivative with respect to time, i. e., $\frac{d}{dt}$.

The following numerical values for the parameters are used [2].

$$\ell = 10^{-2} \text{ sec}; \quad \beta = .0064; \quad \nu = .1 \text{ sec}^{-1}; \quad \alpha_T = 10^{-4} \text{ deg}^{-1};$$
$$(MC_r)_1 = (MC_r)_2 = 100 \text{ kw-sec/deg}; \quad a_1 = a_2 = .01 \text{ sec}^{-1};$$
$$\alpha_{12} = \alpha_{21} = 10^{-2}. \tag{2a}$$

Further, the servomotor constants are assumed to have the values

$$k_1 = k_3 = -20; \quad k_2 = k_4 = .2 \tag{2b}$$

OPTIMAL CONTROL PROBLEM

To obtain the optimal control law for the eighth order nonlinear time delay system of equations (1), it is necessary to deal with a sixteenth order boundary value problem having both advanced and retarded variables. Since the exact solution of this problem is computationally difficult, the method of asymptotic expansions [1, 6] will be used to find an approximation to the exact optimal control. For this purpose a scalar parameter is to be chosen by studying the relative responses of the system variables so as to reduce the number of system equations. Since the power levels n_1 and n_2 are the main variables of interest, the dynamic equations (1a) and (1b) are not neglected. Normally, the variables c_1, c_2, T_1 and T_2 are assumed to be constant by neglecting their dynamic equations, (1c, d, e and f), [2]. Also it can be seen that the coefficients on the right hand side of these equations are small. This indicates that their relative response is slow (large time constants). Similarly in the equation (1g and h), the coefficients that multiply ρ_{e_1} and ρ_{e_2} are large and hence their response is fast (small time constants). Further, in (1a and b) the variables c_1, c_2, T_1 and T_2 are normally neglected since the coefficients that multiply these variables are small. Now all the above coefficients are expressed as linear functions of a scalar parameter λ_2 or its inverse, $1/\lambda_2$, and then the equations (1) are rewritten as

$$\dot{x}_1 = -x_1 + x_2(t-\lambda_1) + \lambda_2(-12.8x_1 + 2y_1) - .2\lambda_2 x_1 y_3 + x_1 z_1 \tag{3a}$$

$$\dot{x}_2 = -x_2 + x_1(t-\lambda_1) + \lambda_2(-12.8x_2 + 2y_2) - .2\lambda_2 x_2 y_4 + x_2 z_2 \tag{3b}$$

$$\dot{y}_1 = \lambda_2(12.8x_1 - 2y_1) \tag{3c}$$

$$\dot{y}_2 = \lambda_2(12.8x_2 - 2y_2) \tag{3d}$$

$$\dot{y}_3 = \lambda_2(.2x_1 - .2y_3) \tag{3e}$$

$$\dot{y}_4 = \lambda_2(.2x_2 - .2y_4) \tag{3f}$$

$$\lambda_2 \dot{z}_1 = -z_1 + u_1 \tag{3g}$$

$$\lambda_2 \dot{z}_2 = -z_2 + u_2 \quad , \tag{3h}$$

where

$x_1(t) = n_1(t)$; $x_2(t) = n_2(t)$; $y_1(t) = c_1(t)$;
$y_2(t) = c_2(t)$; $y_3(t) = T_1(t)$; $y_4(t) = T_2(t)$;
$z_1(t) = 100\,\rho_{e_1}(t)$; $z_2(t) = 100\,\rho_{e_2}(t)$;

and $\lambda_2 = .05$ corresponds to the values given in (2). Here, $\lambda_1 = \tau_{12} = \tau_{21}$ is the neutron transport delay time. This delay time is also normally of the order of .05 seconds [2].

By neglecting the small parameters λ_1 and λ_2, i. e., $\lambda_1 = \lambda_2 = \lambda = 0$, system (3) reduces to

$$\dot{x}_1 = -x_1 + x_2 + x_1 z_1 \tag{4a}$$

$$\dot{x}_2 = -x_2 + x_1 + x_2 z_2 \tag{4b}$$

$$\dot{y}_1 = \dot{y}_2 = \dot{y}_3 = \dot{y}_4 = 0 \tag{4c}$$

$$0 = -z_1 + u_1 \quad ; \quad 0 = -z_2 + u_2 \quad . \tag{4d}$$

This reduced system implies that the delayed neutron power dynamics, the temperature effects and the effect of pure time delay λ_1, are all completely neglected. By using the design of this simple reduced system, an asymptotic series solution for the eighth order system (3) is constructed.

The asymptotic solution is assumed to be of the form [1, 3, 4, 5],

$$\begin{aligned} x(t,\lambda) &= X(t,\lambda) + \lambda m_1(\tau,\lambda) + \lambda n_1(\sigma,\lambda) \\ y(t,\lambda) &= Y(t,\lambda) + \lambda m_2(\tau,\lambda) + \lambda n_2(\sigma,\lambda) \\ z(t,\lambda) &= Z(t,\lambda) + m_3(\tau,\lambda) + n_3(\sigma,\lambda) \end{aligned}$$

where τ and σ are stretched time coordinates,

$$\tau = t/\lambda \ , \qquad \sigma = (T-t)/\lambda \quad .$$

Here the "outer expansion" (X, Y, Z) is the major part of the solution and have an asymptotic power series expansion in λ with coefficients dependent on the slow time variable t, i. e.,

$$(X, Y, Z) \sim \sum_{j=0}^{\infty} \left(X_j(t),\ Y_j(t),\ Z_j(t) \right) \lambda^j \quad .$$

The terms m_i and n_i, i = 1 to 3, represent boundary layer correction terms at the initial point t=0 and the final point t=T respectively. They have power series in λ with coefficients dependent on the stretched variable τ and σ respectively, i. e.,

$$m_i(\tau,\lambda) = \sum_{j=0}^{\infty} m_{ij}(\tau)\lambda^j$$

$$n_i(\sigma,\lambda) = \sum_{j=0}^{\infty} n_{ij}(\sigma)\lambda^j \qquad i = 1 \text{ to } 3 \ .$$

Further, these boundary layer terms will have the following property,

$$m_{ij}(\tau) \to 0 \quad \text{as } \tau \to \infty \ , \qquad n_{ij}(\sigma) \to 0 \quad \text{as } \sigma \to \infty \quad .$$

The asymptotic expansion for the control will be of the form,

$$u = U(t,\lambda) + \nu(\tau,\lambda) + \omega(\sigma,\lambda)$$

where U, ν, and ω all have asymptotic series expansions in λ and the terms ν (or ω) tend to zero as τ (or σ) tends to infinity.

The optimal control problem is to change the power level in each core from a given initial state to the desired final state by minimizing the cost functional

$$J = 5\left(x_1(.5)-x_{1d}\right)^2 + 5\left(x_2(.5)-x_{2d}\right)^2 + .5\int_0^{.5}\left(5(x_1 - x_{1d})^2 + 5(x_2 - x_{2d})^2 + (u_1 - u_{1d})^2 + (u_2 - u_{2d})^2\right)dt \quad .$$

The initial states are

$$x_1(t) = 1 \ ; \quad x_2(t) = 2 \quad \text{for } -\lambda_1 \le t \le 0$$

$$y_1(0) = y_2(0) = 3.2 \ ; \quad y_3(0) = y_4(0) = .5 \ ;$$

$$z_1(0) = z_2(0) = 0 \quad .$$

The final desired states are

$$x_{1d}(.5) = 3 \ ; \quad x_{2d}(.5) = 4 \quad .$$

The values for u_{1d} and u_{2d} are found from the equilibrium conditions, i. e., equating all the derivatives to zero and substituting for x_{1d} and x_{2d},

$$u_{1d} = -1/3 \ ; \quad u_{2d} = 1/4 \quad .$$

NUMERICAL RESULTS

The following three cases are considered:

Case (a): Both singular perturbations and perturbation of time delay: Here a single parameter analysis is done by putting $\lambda_1 = \lambda_2 = \lambda$ in the system of equations (3). In order to use the asymptotic expansion method [1], at first the low-order problem (4) is solved by the method of quasilinearization. Using this low-order design a zero'th-order and first-order asymptotic series approximations of the control are obtained. For comparison purposes an optimal control of the eighth-order time delay system is then calculated. Fig. 3 shows the various controls. It can be noted that the zero'th-order control differs from the nominal control only around the final time t=. 5. There is no significant difference in the performance when these controls are applied to the system (3). Fig. 4 shows the states obtained when different controls are applied to the system (3). It is evident that the first order asymptotic series approximation results in a performance close to the optimal.

Case (b): Singular perturbations only: Here the interest is to study the effect of λ_2 only, i.e., $\lambda_1 = 0$ and $\lambda_2 = \lambda$. The procedure of constructing the asymptotic series is the same as in case (a), except that the time delay parameter is zero in this case. The theoretical aspects of this special case are studied in [4, 5]. Table 1 gives the values of the cost functional with final states in parentheses when the nominal control, the first-order control and the optimal control are applied to the system (3) with $\lambda_1 = 0$.

Case (c): Perturbations of time delay only: Here a second-order time delay system is considered by neglecting delayed neutron and temperature dynamics, i.e., $\lambda_2 = 0$, $\lambda_1 = \lambda$. The method of constructing the asymptotic expansions for this special case is given in [3]. In this case, while finding the optimal control, the nominal control when used as an initial guess did not give convergence whereas the first order control gave convergence in the numerical optimization scheme. Table 1 gives the values of the cost functional with the final states in parentheses when the nominal control, the first order asymptotic control and the optimal control are applied to the system (3) with $\lambda_2 = 0$.

Table 1

Cost Functional J with Final States $x_1(.5)$ and $x_2(.5)$ in Parentheses

Case	Nominal	First-Order	Optimal
a	14.04 (2.22, 2.97)	5.68 (2.78, 3.86)	5.41 (2.96, 3.99)
b	13.10 (2.24, 3.04)	5.47 (2.79, 3.86)	5.30 (3.02, 4.02)
c	4.70 (2.85, 3.88)	4.55 (2.98, 4.02)	4.55 (2.97, 4.00)

It can be noted that the computational difficulties are decreased for cases b and c as compared to the total system (case a). One may try to use the controls obtained in the later two cases as an approximate control to the total system (3). However, they may not be acceptable. Here the performance of the system (3) is seen when the controls obtained in cases b and c are applied. Table 2 gives the values of the cost functional with final states in parentheses. In both cases the cost functional is higher than that of the optimal (5.41) and hence both the singular perturbation as well as the perturbation of time delay play important roles whose effects cannot be neglected in general.

Table 2

Cost Functional J with Final States $x_1(.5)$ and $x_2(.5)$ in Parentheses When First-Order Controls of Case b and c are Applied to the Total System

case b	6.12 (2.68, 3.75)
case c	7.07 (2.54, 3.52)

Fig. 5 compares the computational advantages of the asymptotic method with the second variation method. It can be noted that in the asymptotic method there is a great reduction in the number of system equations to be solved iteratively, thus reducing the computational difficulties.

CONCLUSIONS

The practical applicability of the asymptotic expansions method is shown for a two-core coupled nuclear reactor system. The effect of time delays due to coupling, the effect of neglecting the temperature, and delayed neutron dynamics are studied. It is shown that time delays, temperature and delayed neutron dynamics play a significant role in the control problem. Further, in all the cases considered, the asymptotic method gave results close to the optimal.

REFERENCES

1. Sannuti, P. and Reddy, P. B., "Optimization of High-Order Time Delay Systems by the Method of Asymptotic Expansions," Proceedings of the International Conference on Cybernetics and Society, Washington, D. C., 1972.

2. Weaver, L. E., Reactor Dynamics and Control, Elsevier, New York, 1968.

3. Sannuti, P. and Reddy, P. B., "Asymptotic Series Solution of Optimal Systems with Small Time Delay," IEEE Transactions on AC, vol. AC-18, no. 3, pp. 250-259, June 1973.

4. Sannuti, P., "Asymptotic Expansions of Singularly Perturbed Quasi-linear Optimal Systems," Proc. of Ninth Annual Allerton Conference on Circuit and System Theory, pp. 182-191, 1971.

5. O'Malley, R. E., "On the Asymptotic Solution of Certain Nonlinear Singularly Perturbed Optimal Control Problems," Special Session on Singular Perturbations, JACC 1972.

6. Reddy, P. B., "Optimization of High-Order Time Delay Systems by the Method of Asymptotic Expansions," Ph. D. thesis, Dept. of Electrical Engineering, Rutgers University, 1973.

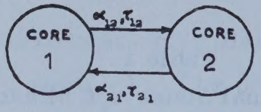

CORE
1
CORE
2
α_{12}, τ_{12}
α_{21}, τ_{21}

Fig. 1 Neutronic coupling

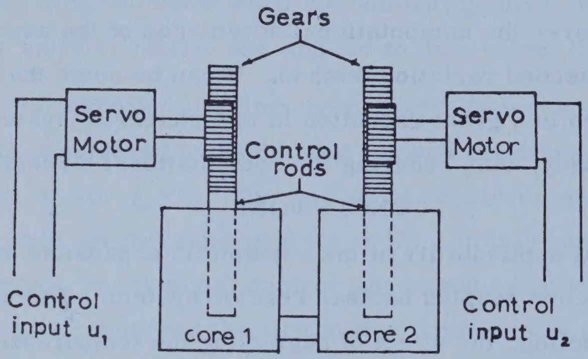

Gears
Servo Motor
Servo Motor
Control rods
Control input u_1
Control input u_2
core 1
core 2

Fig. 2 Coupled-core nuclear reactor system

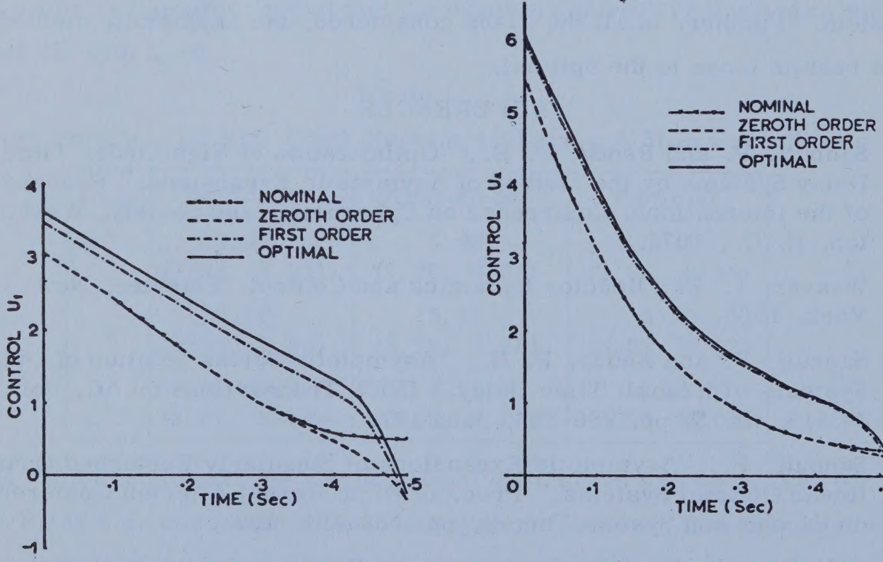


Fig. 3 Comparison of different controls

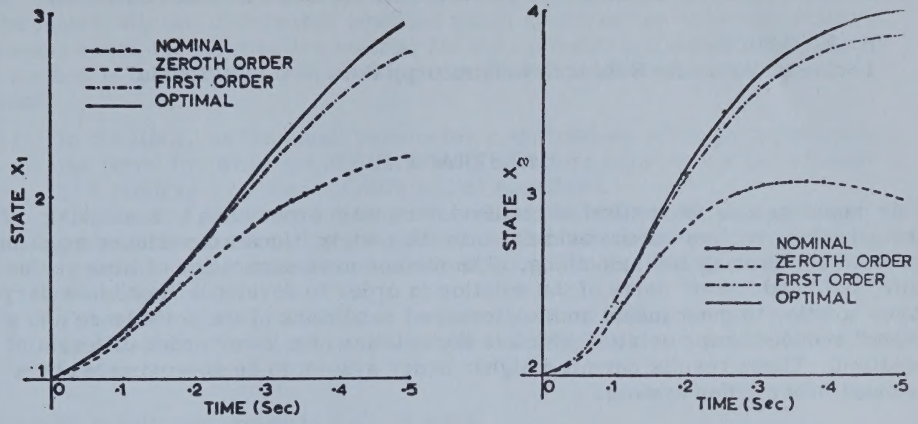


Fig. 4 Behavior of different power levels

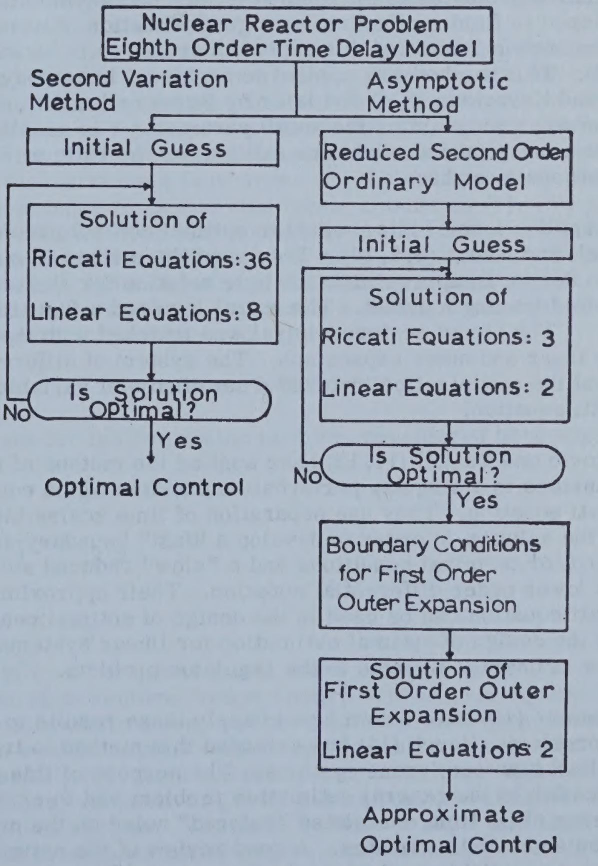

Nuclear Reactor Problem
Eighth Order Time Delay Model
Second Variation Method
Asymptotic Method
Initial Guess
Solution of
Riccati Equations: 36
Linear Equations: 8
Is Solution Optimal?
No
Yes
Optimal Control
Reduced Second Order Ordinary Model
Initial Guess
Solution of
Riccati Equations: 3
Linear Equations: 2
Is Solution Optimal?
No
Yes
Boundary Conditions for First Order Outer Expansion
Solution of First Order Outer Expansion
Linear Equations 2
Approximate Optimal Control

Fig. 5 Comparison of asymptotic and second variation methods

APPLICATION OF SINGULAR PERTURBATION TO OPTIMAL ESTIMATION

H. E. RAUCH
Lockheed Palo Alto Research Laboratory, Palo Alto, California

ABSTRACT

This paper applies the method of matched asmptotic expansions to a singular perturbation problem which could arise in the matrix Riccati covariance equations for optimal filtering and smoothing. The method uses separation of time scales into "fast" and "slow" parts of the solution in order to develop a "fast" boundary-layer solution to meet initial and/or terminal conditions of the covariance and a "slow" reduced order solution which is the solution of a lower order differential equation. These results permit a higher order system to be approximated by a reduced lower order system.

INTRODUCTION

A singular perturbation problem can arise when the highest derivative of a set of differential equations is multiplies by a small parameter ϵ. One way to treat singular perturbation problems is the method of matched asymptotic expansions which was developed in fluid mechanics as a generalization of boundary layer theory. Excellent developments of this approach are given in books by Cole (1) and Van Dyke (2). This method was applied to problems in astrodynamics first by Lagerstrom and Kevorkian (3, 4) and later by Breakwell and Perko (5, 6, 7). In the application to astrodynamics the small parameter ϵ is usually the ratio of planet mass to sun mass (about one to one million) or the ratio of Earth mass to moon mass (about one to eighty).

One of the first applications of this method to optimal control problems was the work of Breakwell and Rauch (8, 9, 10). They used the method of matched asymptotic expansions to derive an approximate analytic solution for the optimal spiral of a low thrust vehicle leaving a planet. The spiral involved a fast time scale of once per revolution. The planet centered spiral was matched with the sun centered trajectory using inner and outer expansions. The system of differential equations was nonlinear and the optimization involved a set of adjoint variables rather than the matrix Ricatti equation.

Recently, Kokotović and Yackel (11, 12) have applied the method of matched asymptotic expansions to a singular perturbation problem which could arise in the matrix Riccati equation. They use separation of time scales into "fast" and "slow" parts of the solution in order to develop a "fast" boundary-layer solution to meet initial and/or terminal conditions and a "slow" reduced solution which is the solution of a lower order differential equation. Their approximate solution to the matrix Riccati equation can be used in the design of optimal control for linear regulators or in the design of optimal estimation for linear systems although their analysis has been primarily directed to the regulator problem.

Haddad and Kokotović (13) have shown how to apply these results to the continuous linear filtering problem. Rauch (14) has extended this method to treating systems with discrete rather than continuous systems. The purpose of this paper is to review the application to the general estimation problem and to extend the previous results to the cases of (1) time correlated "colored" noise on the measurements and (2) calculating the smoothed estimates. A good review of the estimation theory used in this paper is available in the book by Bryson and Ho (15).

For the estimation problem the method of matched asymptotic expansions deals with the matrix Riccati differential equation which specifies the covariance matrix (or in some cases the information matrix) for the optimum estimate. The purpose of the method is to show that the following conditions hold for the covariance equations.

(1) In the limit, as the small parameter ϵ approaches zero, the solution to the (n+m) by (n+m) set of differential equations approaches the solution to a reduced n by n set of differential equations.

(2) For small ϵ, the solution over the entire interval can be approximated with error of order ϵ by the solution to the reduced n by n set plus a boundary-layer correction term to account for the initial conditions of the covariance.

(3) For small ϵ, the solution can be approximated with error of order ϵ^k by a series of k solutions to the reduced set plus a series of k boundary-layer corrections.

The specific results presented in Ref. 14 were:

(a) With a positive definite initial covariance, there is a transformation of variables which allows the covariance equations to be put in a form so that the three conditions, (1), (2), and (3) hold. Using the approximate equations for (2) in the optimal gain introduces additional errors in the covariance of order ϵ^2.

(b) If there is no prior estimate and associated initial covariance, only condition (1) holds for the covariance. However, the information equations (where the information matrix is the inverse of the covariance) can be written so the three conditions (1), (2), and (3) hold.

(c) With discrete rather than continuous measurements, when the time between measurements is of order ϵ^2, the continuous approximation to the covariance equations still meets conditions (1), (2), and (3). When the time between measurements is of order ϵ, the continuous covariance equations hold between measurements and there is a reduced set of discrete update equations at the measurement times. If the time between measurements is not small, a modified reduced set of discrete update equations must be used.

PROBLEM STATEMENT

The system equations for the estimation problem considered here are:

$$dx_1/dt = A_{11}x_1 + A_{12}x_2 + B_1w \tag{1}$$

$$\epsilon dx_2/dt = A_{21}x_1 + A_{22}x_2 + B_xw$$

$$y = C_1x_1 + C_2x_2 + v \tag{2}$$

where t is time with $t_o \leq t \leq t_F$ and ϵ is a small positive scalar, x_1 and x_2 are n- and m- dimensional state vectors, respectively, w is an r-dimensional vector representing system dynamic noise, y is an s- dimensional measurement output vector, and v is an s- dimensional vector representing measurement noise. The symbols A_{ij}, B_i, and C_i represent dynamics matrices, input matrices, and output measurement matrices, respectively, with the appropriate dimensions. All quantities are functions of time. The noise vectors w and v are assumed to have zero mean and their covariance matrices are represented by delta functions.

$$E[v] = E[w] = 0$$

$$\operatorname{cov}[w(t_1)w^T(t_2)] = Q\delta(t_1 - t_2)$$

$$\operatorname{cov}[v(t_1)v^T(t_2)] = R\delta(t_1 - t_2) \tag{3}$$

$$\operatorname{cov}[w(t_1)v^T(t_2)] = S\delta(t_1 - t_2)$$

The initial estimates of the state vectors x_1 and x_2 at the initial time t_o are $\hat{x}_{10}$ and $\hat{x}_{20}$, respectively and the covariance matrices Π_{ij} associated with the initial estimates are known.

$$E[x_i(t_o)] = \hat{x}_{io} \tag{4}$$

$$\operatorname{cov}[x_i(t_o)x_j(t_o)] = \pi_{ij} \tag{5}$$

It is desired to obtain the linear estimate of the state vector which has the minimum mean square error. The well-known solution to the above linear estimation problem is the Kalman-Bucy filter, and involves the solution to the matrix Riccati equation. To illustrate this solution in a form suitable for later analysis it is convenient to define a state vector $\tilde{x}$ composed of vectors x_1 and x_2 and the corresponding set of dynamics, input, and output matrices.

$$\tilde{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ \epsilon^{-1} A_{21} & \epsilon^{-1} A_{22} \end{bmatrix}$$

$$\tilde{B} = \begin{bmatrix} B_1 \\ \epsilon^{-1} B_2 \end{bmatrix} \qquad \tilde{C} = [C_1 \quad C_2] \tag{6}$$

Using the tilde notion the system equations for the estimation problem in Eqs (1) and (2) can be written for $t_o \le t \le t_F$

$$d\tilde{x}/dt = \tilde{A}\,\tilde{x} + \tilde{B}\,w \tag{1a}$$

$$y = \tilde{C}\,\tilde{x} + v \tag{2a}$$

The optimal estimate of the state is $\hat{x}$, the gain on the Kalman-Bucy filter is $\tilde{K}$, and the covariance of the estimate is denoted by $\tilde{P}$. The differential equations for the optimal estimate $\hat{x}$ for $t_o \le t \le t_F$ and the corresponding matrix Ricatti equation for the covariance are

$$d\,\hat{x}/dt = \tilde{A}\,\hat{x} + \tilde{K}[y - \tilde{C}\,\hat{x}] \tag{7}$$

$$\tilde{K} = (\tilde{P}\,\tilde{C}^T + \tilde{B}S)\,R^{-1} \tag{8}$$

$$d\,\tilde{P}/dt = \tilde{A}\,\tilde{P} + \tilde{P}\,\tilde{A}^T - \tilde{K}R\,\tilde{K}^T + \tilde{B}\,Q\,\tilde{B}^T \tag{9}$$

The initial values for the above differential equations are given by Eqs (4) and (5). Notice that when the covariance S is zero, the above equations reduce to the usual filter equations. In order to evaluate the matrix Ricatti Eq (9), , it is convenient to partition the n + m covariance matrix $\tilde{P}$ into submatrices P_{ij} where P_{11} is n by n, P_{12} is n by m and P_{22} is m by m.

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & \epsilon^{-1} P_{22} \end{bmatrix} \quad \text{with} \quad P_{21} = P_{12}^T \tag{10}$$

In terms of these submatrices the matrix Ricatti Eq. (9) can be written

$$dP_{11}/dt = F_{11}(P_{ij}, \epsilon)$$

$$\epsilon dP_{12}/dt = F_{12}(P_{ij}, \epsilon)$$

$$\epsilon dP_{22}/dt = F_{22}(P_{ij}, \epsilon)$$

where
$$\begin{aligned}
F_{11}(P_{ij}, \epsilon) = {} & A^*_{11}P_{11} + A^*_{12}P_{21} + P_{11}A^{*T}_{11} \\
& + P_{12}A^{*T}_{12} - P_{11}W_{11}P_{11} - P_{12}W_{21}P_{11} \\
& - P_{11}W_{12}P_{21} - P_{12}W_{22}P_{22} + Q^*_{11} \\
F_{12}(P_{ij,} \epsilon) = {} & \epsilon A^*_{11}P_{12} + A^*_{12}P_{22} + P_{11}A^{*T}_{21} \\
& + P_{12}A^{*T}_{22} - \epsilon P_{11}W_{11}P_{12} \\
& - \epsilon P_{12}W_{21}P_{12} - P_{11}W_{12}P_{22} \\
& - P_{12}W_{22}P_{22} + Q^*_{12} \\
F_{22}(P_{ij}, \epsilon) = {} & \epsilon A^*_{21}P_{12} + A^*_{22}P_{22} + \epsilon P_{21}A^{*T}_{21} \\
& + P_{22}A^{*T}_{22} - \epsilon^2 P_{21}W_{11}P_{12} \\
& - \epsilon P_{22}W_{21}P_{12} - \epsilon P_{21}W_{12}P_{22} \\
& - P_{22}W_{22}P_{22} + Q^*_{22}
\end{aligned}$$

and where
$$W_{ij} = C_i^T R^{-1} C_j \,, \tag{11}$$

$$A^*_{ij} = A_{ij} - B_i S R^{-1} C_j \;, \quad Q^*_{ij} = B_i (Q - SR^{-1}S^T)\, B_j^T$$

3. PROBLEM SOLUTION

Thus far there have been no approximations and the covariance matrix Ricatti equations in (11) are exact. This section shows how to approximate these exact equations by a reduced set of equations. The form of the covariance Ricatti Eq. (11) is essentially the same as the form of the regular Ricatti equations of Kokotovic and Yackel [11, 12]. Hence, following the duality between estimation and control, the results presented here are similar to the results of Kokotovic and Yackel with an interchange between initial and final boundary conditions.

Notice that the last two differential equations in Eq. (11) have a small parameter ϵ multiplying the left-hand side. Dividing both sides of the equation by ϵ would leave the derivative on the left-hand side and the large quantity ϵ^{-1} multiplying the right-hand side. This means the last two differential equations would respond quite rapidly to perturbations. Thus, the overall solution could be composed of a fast part to meet initial and/or final conditions plus a slow part independent of boundary conditions. One method of solution based on this assumption is called the method of two time scales or, in a more general sense, the method of two variable expansion. The method involves matched asymptotic expansions because both the fast part of the solution and the slow part are approximated by series expansions in the small parameter ϵ.

More explicitly, the partitioned covariance matrices P_{ij} are represented by the sum of a fast part G_{ij} and a slow part H_{ij}. The fast part is a function of initial conditions and normalized time $\tau = t/\epsilon$ and dies out relatively rapidly while the slow part is a function of regular time t.

$$P_{ij} = P_{ij}(t, \epsilon) = \widetilde{G}_{ij}(\tau, \epsilon) + \widetilde{H}_{ij}(t, \epsilon) \quad \text{where} \quad \tau = t/\epsilon \tag{12}$$

The derivative with respect to time t is written

$$dP_{ij}/dt = \epsilon^{-1}\, d\widetilde{G}_{ij}(\tau, \epsilon)/d\tau + d\widetilde{H}_{ij}(t, \epsilon)/dt \tag{13}$$

The fast and slow parts can be represented as an asymptotic series expansion in ϵ. If the matrices A_{ij}, B_i and C_i are functions of ϵ, they should be represented by a similar series. The method of matched asymptotic expansions is implemented by substituting the asymptotic series and Eqs. (12) and (13) into Eq. (11) and collecting terms involving similar powers of ϵ. The terms involving H_{ij} only are used to determine the slow part of the solution while the rest of the terms are used to determine the fast part of the solution.

This paper treats only the first order terms which will be designated H_{ij} and G_{ij}. There are three differential equations from (11) for both the slow and fast parts of the solution. These equations are given below where only the lowest order terms are kept and where the quantities F_{ij} are defined in Eq. (11).

$$dH_{11}/dt + \epsilon^{-1}\, dG_{11}/d\tau = F_{11}(H_{ij} + G_{ij}, 0) \tag{14}$$

$$\epsilon dH_{\ell 2}/dt + dG_{\ell 2}/d\tau = F_{\ell 2}(H_{ij} + G_{ij}, 0) \quad \text{where } \ell = 1, 2 \tag{15}$$

Notice in the first differential equation the large quantity ϵ^{-1} multiplying the derivative $dG_{11}/d\tau$. Thus, in order to satisfy the equation, the derivative of G_{11} must be identically zero. The above equations are satisfied to first order by the following differential equations.

$$dH_{11}/dt = F_{11}(H_{ij}, 0) \tag{16a}$$

$$0 = F_{\ell 2}(H_{ij}, 0) \quad \text{where } \ell = 1, 2 \qquad \text{(16b), (16c)}$$

$$dG_{11}/d\tau = 0$$

$$\begin{aligned} dG_{\ell 2}/d\tau &= F_{\ell 2}(H_{ij} + G_{ij}, 0) - F_{\ell 2}(H_{ij}, 0) \\ &= F_{\ell 2}(H_{ij} + G_{ij}, 0) \text{ where } \ell = 1, 2 \text{ and } \tau = t/\epsilon \end{aligned} \tag{17}$$

In solving these equations (first) Eq. (16c) is solved for H_{22}; (second) Eq. (16b) is solved for H_{12} in terms of H_{11} and H_{22}; and (third) Eq. (16a) is integrated using the values of H_{12} and H_{22} obtained from Eqs. (16b) and (16c). The solution to Eq. (16c) may not be unique because it has several roots. It has been shown in Ref. 11 that under certain conditions a unique positive definite root H_{22} of Eq. (16c) exists such that the root H_{12} of Eq. (16b) is uniquely defined.

When the slow solution quantities H_{ij} are known, the two differential equations (17) for the fast solution quantities G_{12} and G_{22} can be integrated directly. The quantity G_{11} is constant and requires no integration. Since the fast solution must die out, this means that G_{11} is identically zero. The quantities G_{12} and G_{22} need only be calculated for a short time until the effects of initial conditions die out (which means the quantities G_{ij} become of order ϵ).

The initial conditions at t_o for the sum $H_{ij} + G_{ij}$ must satisfy the initial conditions for the covariance in Eq. (5). The slow solution H_{22} will be uniquely determined by Eq. (16c). Likewise, when H_{11} and H_{22} are known, the slow solution H_{12} will be uniquely determined by Eq. (16d). Hence, the fast solutions G_{12} and G_{22} must meet the initial conditions. Because only terms of order ϵ^o are considered, the quantity $\epsilon\Pi_{22}$ can be treated as nil when compared with ϵ^o if desired.

$$\left.\begin{aligned} H_{11} &= \Pi_{11} \\ G_{12} &= \Pi_{12} - H_{12} \\ G_{22} &= \epsilon\Pi_{22} - H_{22} \end{aligned}\right\} \quad \text{at } t = \tau = 0$$

$$\text{and} \quad G_{11} = 0 \text{ for all } t \tag{18}$$

At this point it is appropriate to outline briefly the conditions under which the reduced solution in Eqs. (16), (17), (18) is valid. The following theorem has been proved by Kokotovic and Yackel in Refs. 11 and 12 and is given here without proof.

Theorem: Let the following conditions be satisfied: (a) A_{ij}, B_i, C_i are continuous in t and ϵ; (b) R and Q are positive definite and continuous in t and ϵ; (c) $\widetilde{\Pi}$ is time invariant positive semi-definite and continuous in ϵ; (d) the problem is boundary layer controllable, that is for each fixed $t\epsilon[t_o, t_f]$

$$\text{rank } [C_2^T, (A_{22}^{*T}) C_2^T, \dots, (A_{22}^{*T})^{m-1} C_2^T] = m$$

(e) the problem is boundary layer observable, that is for each fixed $t\epsilon[t_o, t_f]$

$$\text{rank } [B^{*T}, A_{22}^{*} B^{*T}, \dots, A_{22}^{*\,m-1} B^{*T}] = m$$

where B^* is a solution of $B^{*T} B^* = Q_{22}^*$.

Then there exists an $\bar{\epsilon}$ greater than zero such that for all ϵ belonging to $[0, \bar{\epsilon}]$ and all t belonging to $[t_o, t_f]$, the exact solution $P(t, \epsilon)$ of Eqs. (9) through (11) is approximated by

$$P_{ij}(t, \epsilon) = H_{ij}(t) + G_{ij}(t) + 0\ (\epsilon)$$

where $\tau = t/\epsilon$ and a matrix function $f(t, \epsilon)$ is $0(\epsilon)$ if there exist positive constants α and $\bar{\epsilon}$ such that the norm of $f(t, \epsilon)$ satisfies $/f(t, \epsilon)/ \leq \alpha\, \epsilon$ for all ϵ belonging to $[0, \bar{\epsilon}]$. As part of the proof it is shown that there exists for each t a unique positive definite root $H_{22}(t)$ of Eq. (16c) and this root gives a unique solution for $H_{12}(t)$ in Eq. (16b).

4. TIME CORRELATED MEASUREMENT NOISE

An interesting modification of the previous results comes about when there is no longer uncorrelated "nearly white" noise on the continuous measurements. Instead, the measurements are contaminated with time correlated "colored" noise. The general solution to this estimation problem is presented in Section 13.6 of Reference 15. When there is no "nearly white" noise on the continuous measurements, the vector noise v in Eq. (2a) is identically zero.

$$y = \tilde{C}\, \tilde{x} \tag{19}$$

The case of time correlated "colored" noise added to the measurement is included in the above formulation by augmenting the original state vector with the variables in the time correlated noise. The measurement y is modified by taking the derivative with respect to time (which is denoted by superscript dot as in $\dot{y}$) and by multiplying by the small parameter ϵ.

$$\epsilon\, \dot{y} = \epsilon\, \dot{\tilde{C}}\, \tilde{x} + \epsilon\, \tilde{C}\, \dot{\tilde{x}} \tag{20}$$

After making use of Eqs. (1a) and (6) this can be written

$$\epsilon\, \dot{y} = C_1^{*}\, x_1 + C_2^{*}\, x_2 + v^{*}$$

$$\text{where}\quad C_i^{*} = C_2\, A_{2i} + \epsilon\, C_1\, A_{1i} + \epsilon\, \dot{C}_i$$

$$\text{and}\quad v^{*} = (C_2\, B_2 + \epsilon\, C_1\, B_1)\, w \tag{21}$$

Ths quantity $\epsilon\, \dot{y}$ is the "adjusted" measurement and the quantity v^{*} is the "nearly white" noise. It should be pointed out that if the matrix C_2 is identically zero, it is not necessary to multiply the measurement $\dot{y}$ by the small parameter ϵ. The covariance equations for the above estimation problem are identical to those in Eq. (11) except for the changes shown below.

$$C_i \text{ becomes } C_i^{*}$$

$$S = Q\, (C_2\, B_2 + \epsilon\, C_1\, B_1)^T$$

$$R = (C_2\, B_2 + \epsilon\, C_1\, B_1)\, S \tag{22}$$

Thus, when there is time correlated "colored" noise, the singularly perturbed estimation problem can be formulated so all the advantages of order reduction hold. In the development of the covariance equations, it was convenient to take the derivative of the measurement y. However, in the actual implementation of estimation equations similar to Eq. (7), a new state variable x^{*} can be defined in terms of the Kalman-Bucy gain $\tilde{K}$ in Eq. (8), so that it is not necessary to take the derivative of the measurement.

$$x^{*} = \hat{x} - \tilde{K}\, \epsilon\, y$$

$$dx^{*}/dt = (\tilde{A} - \tilde{K}\, C^{*})\, \hat{x} - \tilde{K}\, \epsilon\, \dot{y}$$

where $C^* = (C_1^*, C_2^*)$ (23)

A further modification can arise when there are measurements y_1 with "nearly white" noise and measurements y_2 with time correlated "colored" noise. In that case the measurement y_2 is changed to $\epsilon dy_2/dt$ and the analysis proceeds as before.

5. INFORMATION MATRIX

An alternative formulation of the optimal estimation problem is in terms of the information matrix $\tilde{M}$ which is the inverse of the covariance $\tilde{P}$. The differential equations for a modified estimation variable $\hat{z}$ and the information $\tilde{M}$ can be written directly after substitution into Eqs. (7), (8), and (9).

$$d\hat{z}/dt = -\tilde{A}^T\hat{z} - M\tilde{B}Q\tilde{B}^T\hat{z} + \tilde{C}^T R^{-1} y \tag{24}$$

$$\begin{aligned} d\tilde{M}/dt = & -\tilde{M}\tilde{A} - \tilde{A}^T\tilde{M} + \tilde{M}\tilde{K}R\tilde{K}^T\tilde{M} \\ & -\tilde{M}\tilde{B}Q\tilde{B}^T\tilde{M} \end{aligned}$$

where $$\tilde{M} = \tilde{P}^{-1} \text{ and } \hat{z} = \tilde{M}\hat{x} \tag{25}$$

The required partition of the information matrix $\tilde{M}$ and the initial conditions are given below.

$$\tilde{M} = \begin{bmatrix} M_{11} & M_{21}^T \\ M_{21} & \epsilon M_{22} \end{bmatrix} \tag{26}$$

$$\tilde{M}(t_o) = \begin{bmatrix} \bar{\Pi}_{11} & \bar{\Pi}_{21}^T \\ \bar{\Pi}_{21} & \epsilon\bar{\Pi}_{22} \end{bmatrix} \tag{27}$$

The information Eq. (25) is, in some sense, the dual of the covariance Eq. (9). Because of this dual property, the singular information equation is essentially identical to the singular matrix Ricatti equation which arises in optimal control [11, 12]. In particular, the matrix Ricatti equation for the partitioned matrices M_{ij} is identical to Eq. (11) with the following changes:

$$\begin{aligned} P_{ij} &\rightarrow M_{ij} \\ A_{ij}^* &\rightarrow -A_{ji}^{*T} \\ Q_{ij}^* &\rightarrow W_{ij} \\ W_{ij} &\rightarrow Q_{ij}^* \end{aligned} \tag{28}$$

One appropriate use for the information formulation arises when there is no initial covariance. This means the initial information in Eq. (27) is identically zero with $\Pi_{ij} = 0$. The initial value for the modified estimation variable $\hat{z}(t_o)$ is also zero. It should be pointed out that when the initial information is zero, the covariance formulation is unsatisfactory because there is, in a sense, infinite initial covariance, but the information formulation is still satisfactory.

Another appropriate use for the information formulation is when the Fraser two filter approach is used for smoothing. The covariance Eqs. (7), (8) and (9) can be used for the forward part of the filter and the information Eqs. (24) and (25) can be used on the backward part of the filter. Other smoothing algorithms are discussed in the following section.

6. SMOOTHED ESTIMATE

The filtered estimate $\hat{x}(t)$ is based on measurements from time t_o to time t while the smoothed estimate $\hat{x}_s(t)$ is based on measurements from time t_o to time t_f as discussed in Section 13.3 of Ref. 15. When it is desired to do smoothing, relations hold which are similar to those in filtering. The Kalman-Bucy filter Eqs. (7), (8), and (9) are used to obtain the filtered estimate of the state (designated $\hat{x}$) by starting at the initial time t_o and integrating forward to the final time t_f. The smoothed estimate of the state (designated by $\hat{x}_s$) is obtained by starting with the filtered estimate at the final time t_f and integrating backwards to the initial time t_o. The smoothed estimate is shown below where the quantities M_{ij} are from the information matrix in Eq. (26).

$$d\hat{x}_s/dt = \tilde{A}\,\hat{x}_s + \tilde{L}\,(\hat{x}_s - \hat{x}) \qquad (29)$$

$$\tilde{L} = \begin{bmatrix} L_{11} & L_{12} \\ \epsilon^{-1} L_{21} & \epsilon^{-1} L_{22} \end{bmatrix}$$

where $L_{i1} = Q^*_{i1}\,M_{11} + Q^*_{i2}\,M_{21}$

and $$L_{i2} = Q^*_{i2}\,M_{22} + \epsilon\, Q^*_{i1}\,M_{12} \qquad (30)$$

The above equations are for the case where the measurement noise and dynamic noise are uncorrelated so that S in Eqs. (3) and (11) is identically zero. The above equations require recalculating $\hat{x}$ and M_{ij} on the backward sweep or else storing them on the forward sweep. Notice that M_{ij} can be calculated using the singular perturbation techniques of Eqs. (16) and (17). An alternative formulation is obtained by defining a new vector variable w^*.

$$d\hat{x}_s/dt = \tilde{A}\,\hat{x}_s + \tilde{B}\,Q\,\tilde{B}^T\,w^*$$

$$w^* = \tilde{M}\,(\hat{x}_s - \hat{x}) \qquad (31)$$

$$dw^*/dt = -\tilde{A}^T\,w^* + \tilde{C}^T\,R^{-1}\,\tilde{C}\,\hat{x}_s - \tilde{C}^T\,R^{-1}\,y \qquad (32)$$

The alternate formulation has the advantage integrating a vector equation for w^*, but it has the disadvantage that there is an ϵ^{-2} term in the lower right-hand corner of $\tilde{B}\,Q\,\tilde{B}^T$. The smoothed covariance equations $\tilde{P}_s$ for the smoothed estimate are

$$d\tilde{P}_s/dt = (\tilde{A} - \tilde{L})\,\tilde{P}_s + \tilde{P}_s\,(\tilde{A} - \tilde{L})^T - \tilde{B}\,Q\,\tilde{B}^T$$

$$\widetilde{P}_s = \begin{bmatrix} P_{s11} & P_{s12} \\ P_{s12}^T & P_{s22}/\epsilon \end{bmatrix} \qquad (33)$$

Writing the differential equations for P_{sij} gives a result identical to Eq. (11) with the changes shown below.

$$\begin{aligned} S_{ij} &= 0 \\ A_{ij}^* &= A_{ij} - L_{ij} \\ Q_{ij}^* &= - B_i \, Q \, B_j^T \end{aligned} \qquad (34)$$

7. DISCUSSION OF SOLUTION

It is appropriate to discuss briefly the computational advantages which result from the singular perturbation approach presented here. Assume there are n regular variables represented by x_1 and m singular variables represented by x_2 (or ϵx_2). The standard formulation of the estimation Eq. (7) and the covariance Eq. (9) involves (m+n) and (m+n)(m+n+1)/2 integrations, respectively. The approach used here does not change integration of the estimation Eq. (9), but it approximates the covariance Eq. (9) by the sum of a fast part of the solution (which need be integrated for only a short time to compensate for the boundary conditions of the initial covariance) and a slow part of the solution (which must be integrated over the entire interval). For the first order approximate solution in Eqs. (16) and (17) the fast part of the solution involves integrating differential equations for G_{12} and G_{22} which involve (mn) and n(n+1)/2 integrations, respectively. The slow part of the solution involves integrating differential Eq. (16e) for H_{11} which involves m(m+1)/2 integrations. The slow part of the solution also involves solving the n by n steady state matrix Ricatti Eq. (16c) and inverting an n by n matrix or Eq. (16d). For example, if there are five regular variables (m=5) and five perturbed variables (n=5), the estimation Eq. (6) requires integration of ten variables. The complete covariance Eq. (7) requires integration of 55 variables. The sum of the fast and slow parts of the first order approximate solution requires integration of 55 variables while the slow part by itself requires only 15 variables.

8. REFERENCES

1. Cole, Julian O., "Perturbation Methods in Applied Mathematics," Blaisdell Publishing, Waltham, Mass., 1968.

2. Van Dyke, Milton, "Perturbation Methods in Fluid Mechanics," Academic Press, New York, 1969.

3. Lagerstrom, P.A., and Kevorkian, J., "Matched Conic Approximations to the Two Fixed Force Center Problems," The Astronomical Journal, 68, No. 2, March 1968.

4. Lagerstrom, P.A., and Kevorkian, J., "Earth-to-Moon Trajectories in the Restricted Three-Body Problems," Journal de Mecanique 2, No. 2, June 1963.

5. Perko, L.M., "Interplanetary Trajectories in the Restricted Three-

Body Problem," AIAA Journal, 2, No. 12, December 1964.

6. Breakwell, J. V., and Perko. L. M., "Matched Asymptotic Expansions, Patched Conics, and The Computation of Interplanetary Trajectories," presented at the XVth IAF Congress, Athens, Greece, 13-18 September 1965; and at the AIAA Specialists Meeting, Monterey, California, 16-17 September 1965 (AIAA Paper No. 65-689).

7. Perko, L. M., "Asymptotic Matching in the Restricted Three-Body Problem," Doctoral Dissertation, Department Mathematics, Stanford University, Stanford, California; also available as LMSC Technical Report TR-6-75-65-26.

8. Breakwell, J. V., and Rauch, H. E., "Optimum Guidance for a Low Thrust Interplanetary Vehicle," AIAA Journal, Vol. 4, April 1966, pp. 693-704.

9. Breakwell, J. V. and Rauch, H. E., "Asymptotic Matching in Power-Limited Interplanetary Transfers," Space Flight Mechanics Specialist Conference, Vol. II, Science and Technology Series, pp. 281-300.

10. Breakwell, J. V. and Rauch, H. E., "Three Dimensional Guidance for Low Thrust Escape from a Planet," XIX International Astronautical Congress, New York, 1968, pp. 151-161.

11. Kokotović, P. V. and Yackel, R. A., "Singular Perturbation of Linear Regulators: Basic Theorems," IEEE Transactions on Automatic Control, Vol. AC-17, No. 1, Feb 1972, pp. 29-37.

12. Yackel, R. A. and Kokotović, P. V., "A Boundary Layer Method for the Matrix Riccati Equation," IEEE Transactions on Automatic Control, Vol. AC-18, No. 1, Feb 1973, pp. 17-24.

13. Haddad, A. H. and Kokotović, P. V., "On a Singular Perturbation Problem in Linear Filtering Theory," presented at the 1971 Princeton Conference on Information Systems.

14. Rauch, H. E., "Order Reduction in Estimation with Singular Perturbation," presented at the 1973 Symposium on Nonlinear Estimation and Its Applications, San Diego, California, 10-12 September 1973; to be published in the Proceedings of the Symposium.

15. Bryson, Arthur E., Jr., and Ho, Yu-Chi, "Applied Optimal Control," Blaisdell Publishing Co., Waltham, Mass., 1969.

PARALLEL SOLUTION OF SPARSE SIMULTANEOUS LINEAR EQUATIONS

D. A. CALAHAN
The University of Michigan, Ann Arbor, Michigan

ABSTRACT

A decoupling procedure is given that allows multi-row/column parallel reduction of a sparse matrix. The complexity of the computation is given, and an example is presented.

INTRODUCTION

It is well known that N simultaneous linear equations of the form

$$\underset{\sim}{Ax} = \underline{b}$$

can be solved by parallel operations on entire rows of the composite matrix $\underset{\sim}{C} = [\underset{\sim}{A}\ -\underline{b}]$. Attempts to expand this algorithm to equations with sparse coefficient matrices [1] have taken the approach of expanding each row of $\underset{\sim}{C}$ in a full vector with a length determined by the smallest and largest column indexes present in that row. Due to the routing difficulties associated with current parallel architectures such as Illiac IV, this procedure may be reasonable for matrices that are banded. However, for the matrices that occur in description of electronic, electric power, and mechanical systems, the banded assumption is not viable.

The randomness of nonzero positions in such matrices directly conflicts with the structural regularity expected of problems to be solved in parallel. Therefore, the alternative to forcing a regularity by including zero-valued positions as above is to invest effort in preprocessing, to rearrange the structure until some regularity is achieved.

This restructuring is carried out by reordering the equations and variables by interchanging rows and columns of the matrix. Such ordering procedures are common in solving sparse matrices to control the growth of nonzero positions ("fill") as the equations are solved [2]. Regardless of the criterion, (parallelism or fill minimization) this reordering need be performed only once for a given structure of $\underset{\sim}{A}$, since it is structure- (and not data-) dependent. Consequently, it is most useful when the entries of $\underset{\sim}{A}$ and/or $\underline{b}$ frequently change in value, requiring a sequence of solutions, as during a Newton-type iteration procedure.

ORDERING PROCEDURE

The class of problems to be considered are those with a very low row density (typically 2-5 entries/row). This is an essential requirement since the procedure to be proposed exploits the decoupling between equations. In this sense, the procedure is similar to that proposed by George [3] to reduce multiplications and fill in the solution of finite element problems and by Buneman [4] and Stone [5] to solve tridiagonal systems of equations.

The development of the algorithm is assisted by reviewing the parallel solution of a tridiagonal matrix equation [4]

$$\begin{bmatrix} a_{11} & a_{12} & & & & & \\ a_{21} & a_{22} & a_{23} & & & & \\ & a_{32} & a_{33} & a_{34} & & & \\ & & a_{43} & a_{44} & a_{45} & & \\ & & & a_{54} & a_{55} & a_{56} & \\ & & & & a_{65} & a_{66} & a_{67} \\ & & & & & a_{76} & a_{77} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \end{bmatrix}$$

where in general $a_{ij} \neq a_{ji}$, $i \neq j$. The evaluation of the x_i can proceed by triangular factorization of $\underset{\sim}{A}$, followed by a forward and a backward substitution.

Assuming $\underset{\sim}{A}$ is unsymmetric, the factorization

$$\underset{\sim}{A} = \underset{\sim}{L}\underset{\sim}{U}$$

proceeds as

$$\begin{aligned} \ell_{i+1,i} &= a_{i+1,i} \\ u_{i+1,i} &= a_{i,i+1}/\ell_{ii} \\ \ell_{i+1,i+1} &= a_{i+1,i+1} - a_{i,i+1} * a_{i+1,i} \qquad i = 1,2,\ldots,N-1 \end{aligned}$$

where the ℓ's and u's are the entries of $\underset{\sim}{L}$ and $\underset{\sim}{U}$, with $u_{ii} = 1$, and where $N = 7$. This solution appears to be highly serial; however, the odd and even equations can be decoupled by reordering the rows and columns to give

$$\begin{bmatrix} \underset{\sim}{D}_1^{(0)} & \underset{\sim}{A}_{12}^{(0)} & \underset{\sim}{A}_{13}^{(0)} \\ \underset{\sim}{A}_{21}^{(0)} & \underset{\sim}{D}_2^{(0)} & \underset{\sim}{A}_{23}^{(0)} \\ \underset{\sim}{A}_{31}^{(0)} & \underset{\sim}{A}_{32}^{(0)} & \underset{\sim}{D}_3^{(0)} \end{bmatrix} = \left[\begin{array}{cccc|cc|c} a_{11}^{(0)} & & & & a_{12}^{(0)} & & \\ & a_{33}^{(0)} & & & a_{32}^{(0)} & & a_{34}^{(0)} \\ & & a_{55}^{(0)} & & & a_{56}^{(0)} & a_{54}^{(0)} \\ & & & a_{77}^{(0)} & & a_{76}^{(0)} & \\ \hline a_{21}^{(0)} & a_{23}^{(0)} & & & a_{22}^{(0)} & & \\ & & a_{65}^{(0)} & a_{67}^{(0)} & & a_{66}^{(0)} & \\ \hline & a_{43}^{(0)} & a_{45}^{(0)} & & & & a_{44}^{(0)} \end{array}\right]$$

The factorization can now proceed in two parallel steps: (1) inverting the diagonal matrix $D_1^{(0)}$ and multiplying the first 4 rows by $(D_1^{(0)})^{-1}$ to give $[\underset{\sim}{A}_{12}^{(1)} \; \underset{\sim}{A}_{13}^{(1)}]$ and (2) performing the inner products of all terms in the $[\underset{\sim}{A}_{12}^{(1)} \; \underset{\sim}{A}_{13}^{(1)}]$ partition with terms in the $[\underset{\sim}{A}_{21}^{(0)} \; \underset{\sim}{A}_{31}^{(0)}]$ partitions. The result leaves the updated partitions

$$\begin{bmatrix} (\underset{\sim}{D}_1^{(0)})^{-1} & \underset{\sim}{U}_{12} & \underset{\sim}{U}_{13} \\ \underset{\sim}{L}_{21} & \underset{\sim}{D}_2^{(1)} & \underset{\sim}{A}_{23}^{(1)} \\ \underset{\sim}{L}_{31} & \underset{\sim}{A}_{32}^{(1)} & \underset{\sim}{D}_3^{(1)} \end{bmatrix} = \left[\begin{array}{cccc|cc|c} 1/a_{11}^{(0)} & & & & u_{12} & & \\ & 1/a_{33}^{(0)} & & & u_{32} & & u_{34} \\ & & 1/a_{55}^{(0)} & & & u_{56} & u_{54} \\ & & & 1/a_{77}^{(0)} & & u_{76} & \\ \hline \ell_{21} & \ell_{23} & & & a_{22}^{(1)} & & a_{24}^{(1)} \\ & & \ell_{65} & \ell_{67} & & a_{66}^{(1)} & a_{64}^{(1)} \\ \hline & \ell_{43} & \ell_{45} & & a_{42}^{(1)} & a_{46}^{(1)} & a_{44}^{(1)} \end{array}\right]$$

For example, $u_{56} = a_{56}^{(0)}/a_{55}^{(0)}$ and $a_{66}^{(2)} = a_{66}^{(1)} - \ell_{65}u_{56} - \ell_{67}u_{76}$. The above divide/inner-product steps identify a cycle. They can now be repeated on the lower right-hand matrix (shown in dashes above), since $D_2^{(1)}$ is a diagonal matrix.

The parallelizing of the tridiagonal matrix solution as illustrated above shows the process to be representable as a reordering of the equations and variables. This places the solution in terms common to the sparse matrix literature. The above also suggests the following procedural generalization to sparse matrices with irregular structure.

If an ordering can be found such that the matrix can be written

$$\begin{bmatrix} \underset{\sim}{D}_1^{(0)} & \underset{\sim}{A}_{12} \\ \underset{\sim}{A}_{21} & \underset{\sim}{A}_{22} \end{bmatrix} = \left[\begin{array}{cccc|c} a_{11} & 0 & \cdots & 0 & \\ 0 & a_{22} & \cdots & 0 & \\ \vdots & \vdots & & \vdots & \underset{\sim}{A}_{12} \\ 0 & 0 & \cdots & a_{rr} & \\ \hline & \underset{\sim}{A}_{21} & & & \underset{\sim}{A}_{22} \end{array}\right] \tag{1}$$

then the first r equations can be reduced by performing

(1) a parallel divide step $(\underset{\sim}{D}^{(0)})^{-1}$; $\underset{\sim}{A}_{12}' = [\underset{\sim}{D}^{(0)}]^{-1} \underset{\sim}{A}_{12}$

(2) a parallel inner-product step $\underset{\sim}{A}_{22}' = \underset{\sim}{A}_{22} - \underset{\sim}{A}_{21} \underset{\sim}{A}_{12}'$

Next, $\underset{\sim}{A}_{22}'$ is reordered to give

$$\underset{\sim}{A}_{22}' = \begin{bmatrix} D_2^{(1)} & A_{12}' \\ A_{21}' & A_{22}'' \end{bmatrix} \tag{2}$$

where $D_2^{(1)}$ is a diagonal matrix (with possibly only one entry). The divide/inner-product cycle is repeated to give an A_{12}'' and A_{22}''', etc. Thus the ordering of equations is performed a partition at a time, each partition defined as that producing the largest diagonal matrix $\underset{\sim}{D}_i^{(i-1)}$. Put another way, the algorithm seeks to remove explicit coupling between as many equations as possible, leaving only the implicit coupling through the unreduced part of the matrix, as in $\underset{\sim}{A}'_{12}$ and $\underset{\sim}{A}'_{21}$ in (2).

COMPUTATIONAL COMPLEXITY

At the ith cycle, the best decoupling is seemingly the one which maximizes the size of the diagonal matrix $D_i^{(i-1)}$. (In fact, counterexamples exist which show that fewer total parallel steps can be achieved by choosing a smaller $\underset{\sim}{D}_i^{(i-1)}$ initially. However, the fill produced at each parallel cycle makes such "look ahead" impractical to implement.)

When $\underset{\sim}{A}$ is symmetric, then the matrix structure can be represented by an undirected graph. The decoupling to produce the largest diagonal matrix $\underset{\sim}{D}_i^{(i-1)}$ is analogous to finding the maximum number of nonadjacent nodes in the graph [6] [7]. In the case of an unsymmetric matrix, a corresponding analogy exists with a directed graph. At present, the former problem is felt to have an exponential computational growth with N [8] and the latter problem has not been considered in the literature. For these reasons, a heuristic decoupling algorithm was adopted.

DECOUPLING ALGORITHM

The decoupling proceeds by accepting the (1,1) position as the first position in the yet to be produced diagonal matrix $\underset{\sim}{D}_1^{(0)}$. The remaining matrix is then scanned beginning in the upper left-hand corner to find a row r and column s such that

(1) both $a_{r,1}$ and $a_{1,s}$ are zero-valued, and
(2) $a_{r,s}$ is not zero-valued.

The r row and s column are then exchanged with the second row and column, leaving a 2 x 2 diagonal submatrix in the upper left-hand corner. A new search is then initiated to find a row and column that are zero-valued in the first two positions, and nonzero-valued at the row-column intersection. This search-exchange process continues until it becomes impossible to find a row and a column that are zero-valued in the first t positions, and nonzero valued at the row-column intersection. $\underset{\sim}{D}_1^{(0)}$ is then a (t-1) x (t-1) matrix. In a similar way, $\underset{\sim}{D}_i^{(i-1)}$ is found at the ith parallel reduction cycle.

RESULTS

This ordering procedure has been applied to a matrix ($\underset{\sim}{A}_s$) resulting from a simulation of the three-dimensional model of an automotive suspension system. The matrix is of dimension 83 x 83 with an average of 5 entries/row. This is a dense matrix by most standards, so that the decoupling that can be expected is less than for a more sparse matrix. The resultant LU map of nonzero positions is shown in Figure 1. A total of 13 $\underset{\sim}{D}_i^{(i-1)}$ can be identified, meaning that the factorization can be carried out in 12 parallel cycles. The partition sizes are given in Table I.

To study the dependence of the number of parallel cycles on the matrix size and sparsity, a sequence of matrices were generated with randomly-positioned entries. The results are shown in Figure 2(a). As expected, the number of cycles increases with the matrix size and row density. It is interesting to note that $\underset{\sim}{A}_s$ could be factored in somewhat more than 1/2 the number of cycles than a randomly-generated matrix with the same size and same number of entries/row.

Table I. Summary of results of partitioning $\underset{\sim}{A}_s$.

Parallel Cycle No.	Mult. & Div.	No. of Rows
1	116	40
2	132	20
3	135	8
4	85	4
5	12	1
6	11	1
7	65	3
8	25	1
9	21	1
10	13	1
11	7	1
12	3	1
Total	625	83

Another concern in any matrix ordering is the number of fills created. As observed in Figure 2(b) for the random matrices, the fills resulting from parallel ordering are not significantly greater in number than those obtained from an ordering (Markowitz [1]) to reduce the number of fills. The suspension system problem shows a very low fill count when fill minimization is used, the fill count returning to that of a random matrix when parallel ordering is used.

The discrepancies noted above between the random matrices and $\underset{\sim}{A}_s$ possibly arise from the tendency of the latter to have local patterns and even small dense blocks of nonzero positions. This adds a form of regularity to the matrix, possibly aiding both parallelism and fill minimization, depending on the ordering criterion.

ORDER OF CALCULATION

The preceding factorization algorithm, wherein a row and column partition are simultaneously reduced, is analogous to Crout reduction. This procedure is illustrated in detail in Figure 3(a) for a 4 x 4 partition. Important to parallelism is the fact that, as Table I shows, the number of multiplications and divisions can vary considerably from cycle-to-cyle. The first cycles involve large, sparse partitions; the latter cycles involve small, dense partitions. In the final cycles, even a dense full row and column reduction cannot occupy a large number of processors.

An alternative calculation sequence is shown in Figure 3(b), where the $\underset{\sim}{L}$ and $\underset{\sim}{U}$ factors are determined a row at a time, analogously to Gauss elimination. Here, the $\underset{\sim}{L}_{ij}$ and $\underset{\sim}{U}_{ik}$ cannot be determined in parallel across a single row i. For example, calculation of $\underset{\sim}{U}_{ik}$ requires knowledge of $\underset{\sim}{L}_{ij}$, for $j < i$. However, parallelism can still be exploited within a partition. Should this prove inadequate, processing could be initiated on calculation of $\underset{\sim}{L}_{\ell k}$ ($\ell > i$, $k < i$), as specified by the Crout ordering.

PARALLELISM AND $\underset{\sim}{A}^{-1}$

The explicit inverse of a sparse matrix $\underset{\sim}{A}$ is usually far more dense than the matrix $\underset{\sim}{A}$ itself. For example, the inverse of a tridiagonal matrix is a full matrix. Moreover, the number of nonzero positions in $\underset{\sim}{A}^{-1}$ is independent of ordering, since $\underset{\sim}{A}^{-1}$ will have a nonzero position in the (i,j) position if and only if x_i is a function of b_j. It is interesting to note, however, that the $\underset{\sim}{A}^{-1}\underline{b}$ calculation is a single parallel inner-product step. Therefore, in applications where $\underset{\sim}{A}$ is a constant matrix but $\underline{b}$ varies, it may be feasible to consider the calculation and storage of $\underset{\sim}{A}^{-1}$.

A far more realistic alternative follows from the observation that

$$\underset{\sim}{A}^{-1} = \underset{\sim}{U}^{-1}\underset{\sim}{L}^{-1}$$

Clearly, since the density of $\underset{\sim}{L}$ and $\underset{\sim}{U}$ is dependent on the ordering, so too is the density of $\underset{\sim}{L}^{-1}$ and $\underset{\sim}{U}^{-1}$. The calculation $\underline{y} = \underset{\sim}{L}^{-1}\underline{b}$ and $\underline{x} = \underset{\sim}{U}^{-1}\underline{y}$ is each a parallel inner-product step. Therefore, even though $\underset{\sim}{L}^{-1}$ and $\underset{\sim}{U}^{-1}$ are more dense than $\underset{\sim}{L}$ and $\underset{\sim}{U}$ [9] it may be advantageous to store $\underset{\sim}{L}^{-1}$ and $\underset{\sim}{U}^{-1}$ rather than $\underset{\sim}{L}$ and $\underset{\sim}{U}$, when A is constant.

To study the density of $\underset{\sim}{L}^{-1}$ and $\underset{\sim}{U}^{-1}$, the augmented matrix

$$\begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{I} \\ \underset{\sim}{I} & \underset{\sim}{O} \end{bmatrix}$$

can be triangularly factored to yield

$$\begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{I} \\ \underset{\sim}{I} & \underset{\sim}{O} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{L} & \underset{\sim}{O} \\ \underset{\sim}{U}^{-1} & \text{◢} \end{bmatrix} \begin{bmatrix} \underset{\sim}{U} & \underset{\sim}{L}^{-1} \\ \underset{\sim}{O} & \text{◢} \end{bmatrix} \tag{3}$$

where the cross-hatched calculations are not performed. It is a straightforward process to modify an optimal ordering algorithm [2] to suppress the fills occurring in $\underset{\sim}{L}^{-1}$ and $\underset{\sim}{U}^{-1}$ in (3) rather than in $\underset{\sim}{L}$ and $\underset{\sim}{U}$. The result of applying such a minimization to the matrix describing a suspension system $(\underset{\sim}{A}_s)$ results in the ratio

$$\frac{(\#\text{ of nonzero positions in }\underset{\sim}{L}^{-1},\ \underset{\sim}{U}^{-1})}{(\#\text{ of nonzero positions in }\underset{\sim}{L},\ \underset{\sim}{U})} = \frac{912}{508} \cong 1.8$$

This ratio is similar to the factor of 2 obtained for random matrices noted in [9]. These results indicate that for this example it is quite reasonable to store $\underset{\sim}{L}^{-1}$ and $\underset{\sim}{U}^{-1}$ when they are constant.

CONCLUSIONS

Although it is common to associate parallel computation algorithms with large, horizontally-organized machines such as Illiac IV, it should be recognized that any architecture where operands are moved en masse—with a subsequent temporary loss of addressibility—has similar requirements. Pipelined processors—such as the CDC STAR and TI ASC—and minicomputers requiring disc storage for large matrices, could also be benefited by the proposed decomposition.

It is also clear that the above procedure could be applied to block partitions of a matrix rather than single matrix entries. In this sense, it is an algorithmic generalization of George's method [5].

ACKNOWLEDGMENTS

This work was performed under sponsorship of the AFOSR Mathematical and Information Sciences Directorate (Grant No. 71-2027). The author is indebted to Mr. Mike Schlansker and Mr. Tom Grapes for programming assistance in preparing the results.

REFERENCES

[1] Y. MATSUSHITA. Sparse matrix inversion on Illiac IV. Document 140, Center for Advanced Computation, University of Illinois, June 1968.

[2] H. M. MARKOWITZ. The elimination form of the inverse and its application to linear programming. Management Science, Vol. 3, pp. 255-269, 1957.

[3] A. S. GEORGE. Block elimination on finite element systems of equations. Sparse Matrices and Their Applications (Ed. by Rose and Willoughby). Plenum Press, 1972.

[4] OSCAR BUNEMAN. A compact non-iterative poisson solver. Report 294, Stanford University Institute for Plasma Research, 1969.

[5] H. S. STONE. An efficient parallel algorithm for the solution of a tridiagonal linear system of equations. Report SU-SEL-81-063, Stanford Electronics Laboratory, December 1971.

[6] F. HARARY. Graph theory. Addison-Wesley, 1971.

[7] L. E. TROTTER. Solution characteristics and algorithms for the vertex packing problem. Technical Report 168, Department of Operations Research, Cornell University, January 1973.

[8] R. M. KARP. Reducibility among combinatorial problems. Complexity of Computer Computations (Ed. by Miller and Thatcher). Plenum Press, 1972.

[9] R. K. BRAYTON, F. G. GUSTAVSON, and R. A. WILLOUGHBY. Some results on sparse matrices. Proceedings Sparse Matrix Conference, IBM Thomas Watson Research Center, Yorktown, N.Y., 1968.

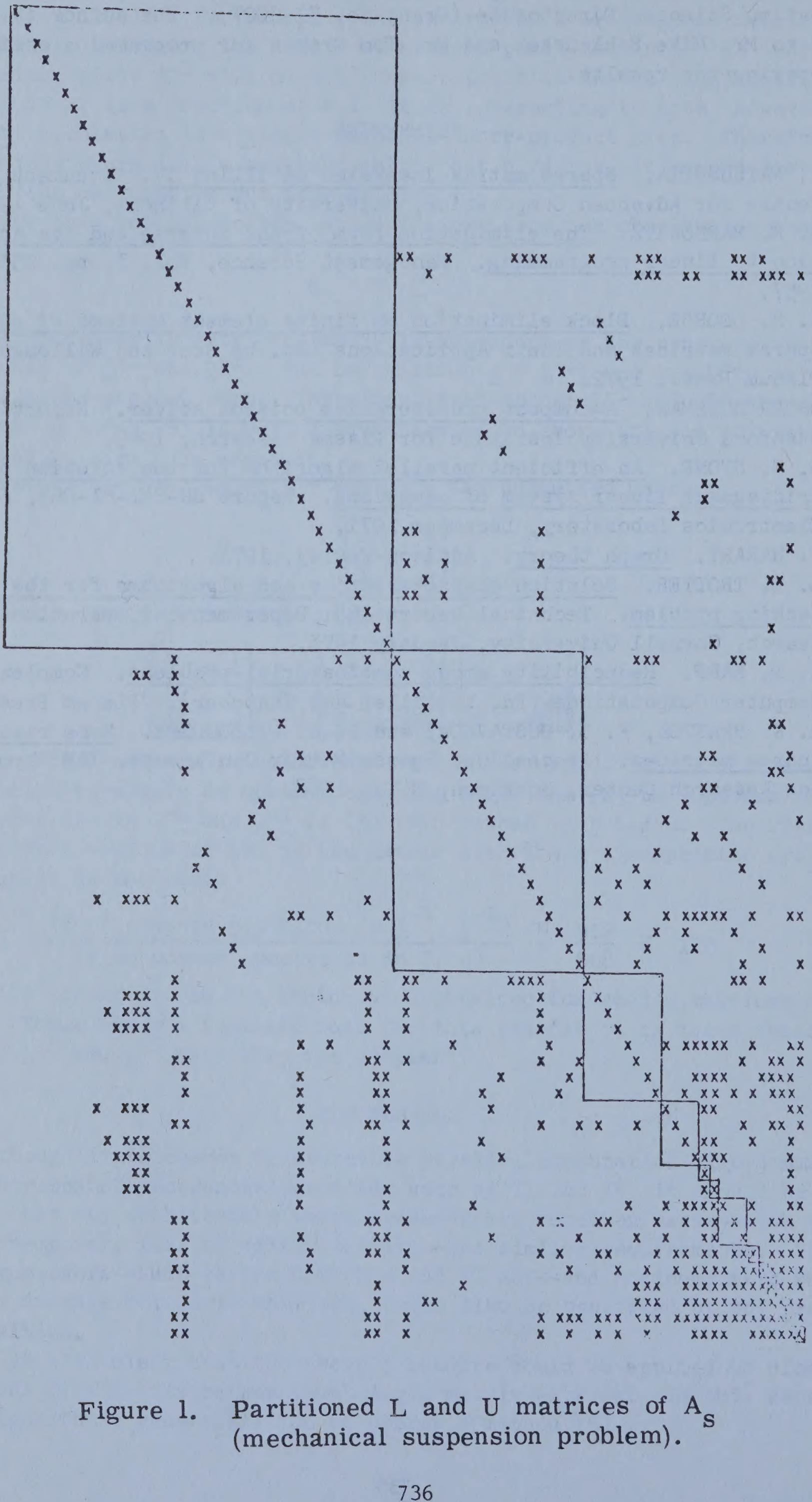


Figure 1. Partitioned L and U matrices of A_S (mechanical suspension problem).

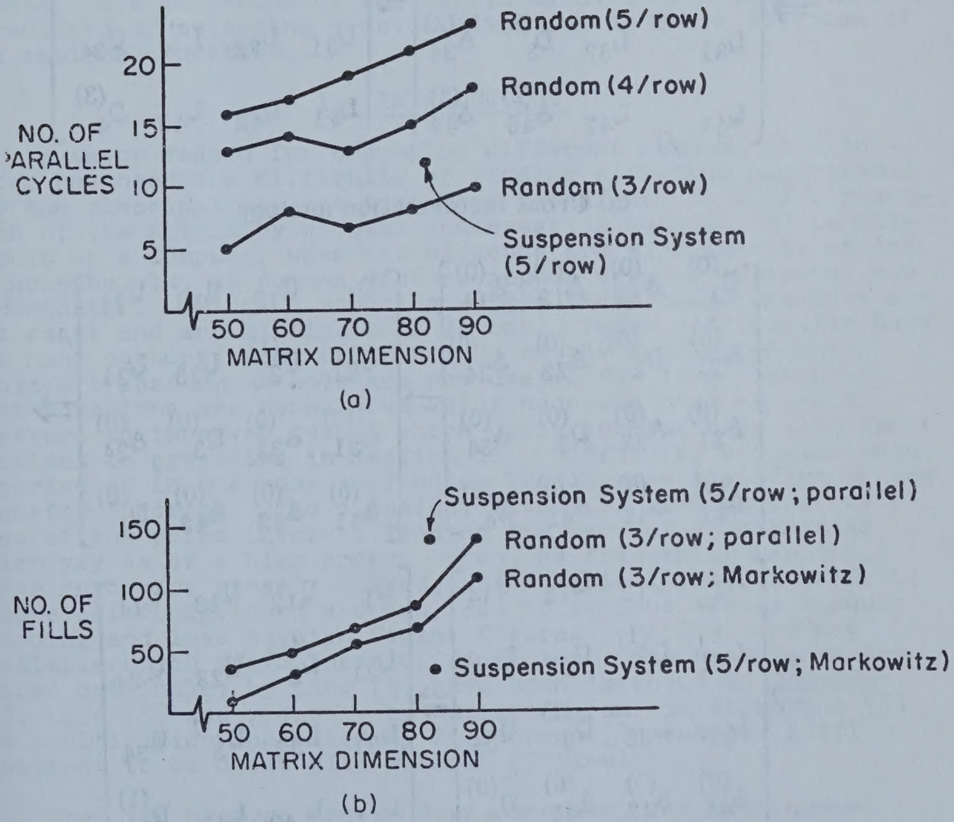


Figure 2. Parallelism and fill studies

$$\begin{bmatrix} D_1^{(0)} & A_{12}^{(0)} & A_{13}^{(0)} & A_{14}^{(0)} \\ A_{21}^{(0)} & D_2^{(0)} & A_{23}^{(0)} & A_{24}^{(0)} \\ A_{31}^{(0)} & A_{32}^{(0)} & D_3^{(0)} & A_{34}^{(0)} \\ A_{41}^{(0)} & A_{42}^{(0)} & A_{43}^{(0)} & D_4^{(0)} \end{bmatrix} \Rightarrow \begin{bmatrix} (D_1^{(0)})^{-1} & U_{12} & U_{13} & U_{14} \\ L_{21} & D_2^{(1)} & A_{23}^{(1)} & A_{24}^{(0)} \\ L_{31} & A_{32}^{(1)} & D_3^{(1)} & A_{34}^{(1)} \\ L_{41} & A_{42}^{(1)} & A_{43}^{(1)} & D_4^{(1)} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \begin{bmatrix} (D_1^{(0)})^{-1} & U_{12} & U_{13} & U_{14} \\ L_{21} & (D_2^{(0)})^{-1} & U_{23} & U_{24} \\ L_{31} & L_{32} & D_3^{(2)} & A_{34}^{(2)} \\ L_{41} & L_{42} & A_{43}^{(2)} & A_{44}^{(2)} \end{bmatrix} \Rightarrow \begin{bmatrix} I & U_{12} & U_{13} & U_{14} \\ L_{21} & I & U_{23} & U_{24} \\ L_{31} & L_{22} & I & U_{34} \\ L_{41} & L_{42} & L_{43} & D_4^{(3)} \end{bmatrix}$$

(a) Crout factorization analogy

$$\begin{bmatrix} D_1^{(0)} & A_{12}^{(0)} & A_{13}^{(0)} & A_{14}^{(0)} \\ A_{21}^{(0)} & D_2^{(0)} & A_{23}^{(0)} & A_{24}^{(0)} \\ A_{31}^{(0)} & A_{32}^{(0)} & D_3^{(0)} & A_{34}^{(0)} \\ A_{41}^{(0)} & A_{42}^{(0)} & A_{43}^{(0)} & D_4^{(0)} \end{bmatrix} \Rightarrow \begin{bmatrix} D_1 & U_{12} & U_{13} & U_{14} \\ L_{21} & D_2^{(1)} & L_{23} & L_{24} \\ A_{31}^{(0)} & A_{32}^{(0)} & D_3^{(0)} & A_{24}^{(0)} \\ A_{41}^{(0)} & A_{42}^{(0)} & A_{43}^{(0)} & D_4^{(0)} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \begin{bmatrix} D_1 & U_{12} & U_{13} & U_{14} \\ L_{21} & D_2^{(1)} & U_{23} & U_{24} \\ L_{31} & L_{32} & D_3^{(1)} & U_{34} \\ A_{41}^{(0)} & A_{42}^{(0)} & A_{43}^{(0)} & D_4^{(0)} \end{bmatrix} \Rightarrow \begin{bmatrix} D_1 & U_{12} & U_{13} & U_{14} \\ L_{21} & D_2^{(1)} & U_{23} & U_{24} \\ L_{31} & L_{32} & D_3^{(1)} & U_{34} \\ L_{41} & L_{42} & L_{43} & D_4^{(1)} \end{bmatrix}$$

(b) Gauss elimination analogy

Figure 3. Alternative factorization procedures.

A SYMBOLIC METHOD TO COMPUTER-AIDED DESIGN OF LINEAR ACTIVE NETWORKS BASED ON A NODAL APPROACH

N. B. Rabbat*
University of Alabama
Department of Electrical Engineering
University, Alabama
and
Washington University
Computer Systems Laboratory
St. Louis, Missouri

ABSTRACT

A new symbolic method to computer-aided design of linear active networks based on a nodal approach is presented. The topological considerations of the method depend on recognizing each term of the admittance matrix in terms of its 'S' coefficients. The inversion of the matrix by pivotal condensation is formulated by selecting a novel pivot for an exact division of the rational functions.

I. INTRODUCTION

A major reason for employing different methods on a computer has been the difficulty of finding efficient algorithms for the classical Laplace transform technique. A direct reduction of the matrix by pivotal condensation methods [1] is difficult on a computer when the elements are expressed as ratios of polynomials, as common factors appear between numerator and denominator. Owing to rounding errors these common factors are not exact and are difficult to detect. These difficulties have led many authors, e.g. Calahan [2], to the conclusion that Laplace transform methods are not useful for large networks when equations are formulated using node and mesh analysis. However, an improved method which deals successfully with these problems is presented in Section II. Difficulty has also been experienced in the past in finding the inverse transform of the transfer function. The classical method of summing the residues at the poles involves finding the zero of a polynomial which may be of a high order. When, as frequently happens, these zeros are closely spaced, the numerical process tends to become ill-conditioned and can lead to serious errors through rounding and loss of significant figures. Methods are now available which avoid these difficulties. The techniques described originally by Liou [3] have been improved by Thomson [4], Liou [5] and recently by Muraleedharan and Stromsmoe [6]. The modification suggested by Silverberg [7] enables stiff equations to be dealt with without difficulty.

The Liou process breaks down, however, for high order systems. A new method [8] for the inverse Laplace transform

*The research presented in this paper was supported by the Advanced Research Projects Agency of the Department of Defense under Contract SD-302 and by the Division of Research Facilities and Resources under NIH grants FR-00218 and RR-00396.

which uses the FFT (Fast Fourier Transform), modified weight factors and the shifting theorem to deal successfully with high order systems makes the formulation of the circuit as described in Section II by symbolic methods [9-11] even more attractive.

II. NODAL MATRICES IN SYMBOLIC FORM

The presentation will not be complicated at this stage by explaining the method for constructing the nodal admittance matrix in symbolic form and the algorithmic features associated with it -- a section which is certainly interesting and original. It should be pointed out, however, that the features of the procedure of setting the nodal admittance matrix in symbolic form do not result in more computation as previously stated [24]. We will now assume that the nodal admittance matrix has been set up and the distinction between the terms of the rational functions has been made; and we will proceed to discuss the method of inversion of the nodal matrices. Suppose we have the admittance matrix

$$\psi(n) = \{\psi_{ij}\} = \frac{\alpha_{ij}}{\beta^{(0)}} \quad , \; i,j = 1,2,\ldots,n \tag{1}$$

where the terms are normalized in such a way that the common denominator $\beta^{(0)}$ is either 1 or s. Using a pivotal condensation procedure and selecting ψ_{nn} as the first pivot, we can replace each term ψ_{ij}, $i,j < n$, by $\psi_{ij}^{(1)}$ where

$$\psi_{ij}^{(1)} = \psi_{ij} - \frac{\psi_{in}\,\psi_{ij}}{\psi_{nn}} \tag{2}$$

$$= \frac{\alpha_{ij}\,\alpha_{nn} - \alpha_{in}\,\alpha_{nj}}{\beta^{(0)}\,\alpha_{nn}} = \frac{\alpha_{ij}^{(1)}}{\beta^{(0)}\,\alpha_{nn}} \tag{3}$$

where obvious cancellations of $\beta^{(0)}$ have been made. If we now pivot about $\psi_{n-1,n-1}^{(1)}$, then for $i,j < n-1$, $\psi_{ij}^{(1)}$ becomes

$$\psi_{ij}^{(2)} = \frac{\alpha_{ij}^{(1)}}{\beta^{(0)}\,\alpha_{nn}} - \frac{\alpha_{i,n-1}^{(1)}\,\alpha_{n-1,j}}{\beta^{(0)}\,\alpha_{n-1,n-1}^{(1)}} \tag{4}$$

If (4) is evaluated in terms of unprimed quantities using (3), the common factor α_{nn} appears in both numerator and denominator. The same type of cancellation occurs in subsequent states, and if the original matrix $\psi_{(n)}$ was full it would be simple to predict the occurrence of these common factors. In practice, however, $\psi_{(n)}$ is generally a sparse matrix and the common factors are more difficult to predict as their positions no longer follow such a simple law. Since, owing to rounding errors in the computer, these common factors are not exact, they can only be removed in practice if their occurrence can be predicted. This greatly complicates the method of obtaining the inverse. This new method avoids these difficulties. Introduced originally by Shipley and Coleman [10] as a method for inverting numerical matrices it is developed as an efficient technique for inverting matrices of circuit analysis whose elements are expressed in terms of s.

We will rewrite (2) as follows:

$$\psi_{ij}^{(n-k+1)} = \frac{\psi_{ij}^{(n-k)}\psi_{kk}^{(n-k)} - \psi_{ik}^{(n-k)}\psi_{kj}^{(n-k)}}{\psi_{kk}^{(n-k)}} \tag{5}$$

where kk is the pivotal term operating on term ij (k,j $\neq$ k), and the superscript indicates the number of iterations which have taken place on a particular term. Clearly all terms involved in the next iteration will have $\psi_{kk}^{(n-k)}$ as a divisor, and it can be seen that, in an operation in the form of (5), this term will cancel once and still remain in the first power of the denominator. Thus division by a pivotal term may be carried out on the iteration immediately following its use as pivot. This (5) may be altered to read

$$\psi_{ij}^{(n-k+1)} = \frac{\psi_{ij}^{(n-k)}\psi_{kk}^{(n-k)} - \psi_{ik}^{(n-k)}\psi_{kj}^{(n-k)}}{\psi_{k+1,k+1}^{(n-k-1)}} \tag{6}$$

where $\psi_{k+1,k+1}^{(n-k-1)}$ is the pivot employed on the iteration immediately preceding the one using $\psi_{kk}^{(n-k)}$.

It can be proved that $\psi_{11}^{(n-1)}$ will be the determinant of the original array. It is also true that the division involved in (6) will always take place exactly, i.e., without remainder. This is shown by considering the term ij, where i,j < n-1. The first iteration gives

$$\psi_{ij}^{(1)} = \psi_{ij}\psi_{nn} - \psi_{in}\psi_{nj} \tag{7}$$

the previous pivot being considered equal to unity. The next iteration will then give

$$\psi_{ij}^{(2)} = \frac{\psi_{ij}^{(1)}\psi_{n-1,n-1}^{(1)} - \psi_{i,n-1}^{(1)}\psi_{n-1,j}^{(1)}}{\psi_{nn}} \tag{8}$$

Substitution of terms in the form of (7) into (8) shows that terms which are not factors of ψ_{nn} in the numerator of (8) disappear, so that the division takes place exactly. Since the numerator of (8) is exactly of the same form as (7), it is clear that, for i,j < n-2, division by $\psi_{n-1,n-1}^{(1)}$ will take place exactly after the third iteration. The inversion rules will be modified here in the light of our previous discussion. If ψ_{ij} is the general element in the admittance matrix $\psi_{(n)}$, and ψ_{kk} is the diagonal element selected for pivot, then the rules are as follows: (i) All elements external to the row and column containing the diagonal element are modified so that ψ_{ij} becomes

$$\frac{1}{\psi_{k+1,k+1}^{(1)}}\{\psi_{ik}\psi_{kk} - \psi_{ik}\psi_{kj}\} \tag{9}$$

where $\psi_{k+1,k+1}^{(1)}$ is the previous diagonal pivot. When ψ_{kk} is the first pivot selected then the previous diagonal pivot is taken to be unity. (ii) The element ψ_{kk} is replaced by $-\psi^{(1)}_{k+1,k+1}$ (or by -1 when ψ_{kk} is the first pivot). Each other element in the row and column to ψ_{kk} is multiplied by -1. (iii) After the last diagonal operation, the denominator of each term must be multiplied by the last pivot selected.

The extension of this method to rational functions is important and we will proceed as follows. Now, considering rational terms of the nodal-admittance matrix, with numerator and denominator polynomials, the first iteration gives

$$\frac{\alpha^{(1)}_{ij}}{\beta^{(1)}_{ij}} = \frac{\alpha_{ij}\alpha_{nn}\beta_{in}\beta_{nj} - \alpha_{in}\alpha_{nj}\beta_{ij}\beta_{nn}}{\beta_{in}\beta_{nj}\beta_{ij}\beta_{nn}} \qquad (10)$$

Equation (10) may be written

$$\frac{\alpha^{(1)}_{ij}}{\beta^{(1)}_{ij}} = \frac{\alpha_{ij}\alpha_{nn} - \alpha_{in}\alpha_{nj}\,\dfrac{\beta_{ij}\beta_{nn}}{\beta_{in}\beta_{nj}}}{\beta_{ij}\beta_{nn}} \qquad (11)$$

For networks with controlled sources which are not frequency dependent, the division of terms in the numerator of (11) will always take place without remainder (i.e., $\beta_{in}\beta_{nj}$ is a factor of β_{nn}).

Equation (11) may be written

$$\frac{\alpha^{(1)}_{ij}}{\beta^{(1)}_{ij}} = \frac{\alpha^{(1)}_{ij}}{\beta_{ij}\beta_{nn}} \qquad (12)$$

On the next iteration we get, after division by the previous pivot

$$\frac{\alpha^{(1)}_{ij}}{\beta^{(1)}_{ij}} = \frac{\dfrac{1}{\alpha_{nn}}\left\{\alpha^{(1)}_{ij}\,\alpha^{(1)}_{n-1,n-1} - \alpha^{(1)}_{i,n-1}\alpha^{(1)}_{n-1,j}\,\dfrac{\beta_{ij}\beta_{n-1,n-1}}{\beta_{i,n-1}\beta_{n-1,j}}\right\}}{\beta_{ij}\beta_{n-1,n-1}\beta_{nn}} \qquad (13)$$

Assuming that ψ_{ij} can be recognized in terms of its "a" coefficients, and can be written as follows

$$\psi_{ij} = \frac{a_{ij(2)}s^2 + a_{ij(1)}s + a_{ij(0)}}{\beta_{ij}\begin{Bmatrix}1\\ s\end{Bmatrix}} \qquad (14)$$

we can write for the element ψ_{ij}, $i < k$, $j < k$,

$$\psi_{ij} = \frac{1}{\psi'_{k+1,k+1}}\left\{\frac{\left[\sum_{i=0}^{j} a_{ij(n-k-1)}a_{kk(n-k-j+1)}s^{n-k(n-k-j)}\begin{Bmatrix}s^2\\ s\\ 1\end{Bmatrix}\right] - \left[\sum_{i=0}^{j} a_{ik(n-k-1)}a_{kj(n-k-j+1)}s^{n-k(n-k-j)}\begin{Bmatrix}s^2\\ s\\ 1\end{Bmatrix}\right]}{\beta_{ij}\prod_{x=k}^{n}\prod_{y=k}^{n}\beta_{xy}\prod_{y=k}^{n}\beta_{iy}\prod_{x=k}^{n}\beta_{xj}}\right\} \qquad (15)$$

For the element ψ_{ij}, $i > k$, $j < k$, ψ_{ij}, similar to (15), becomes

$$\psi_{ij} = \frac{1}{\alpha'_{k+1,k+1}} \frac{\left\{ \left[\sum_{i=0}^{j} a_{ij(n-k-1)} a_{kk(n-k-j+1)} s^{n-k(n-k-j)} \left\{ \begin{matrix} 1 \\ s \end{matrix} \right\} \right] - \left[\sum_{i=0}^{j} a_{ik(n-k-1)} a_{kj(n-k-j+1)} s^{n(n-j)} \left\{ \begin{matrix} 1 \\ s \end{matrix} \right\} \right] \right\}}{\prod_{x=k}^{n} \beta_{xj} \prod_{x=k}^{n} \prod_{y=k}^{n} \beta_{xy} \prod_{x=i}^{n} \beta_{xi}} \qquad (16)$$

For the element ψ_{ij}, $i < k$, $j > k$, ψ_{ij} becomes

$$\psi_{ij} = \frac{1}{\alpha'_{k+1,k+1}} \frac{\left\{ \left[\sum_{i=0}^{j} a_{ij(n-k-1)} a_{kk(n-k-j+1)} s^{n-k(n-k-j)} \left\{ \begin{matrix} 1 \\ s \end{matrix} \right\} \right] - \left[\sum_{i=0}^{j} a_{ik(n-k-1)} a_{kj(n-k-j+1)} s^{n-k(n-k-j)} \left\{ \begin{matrix} 1 \\ s \end{matrix} \right\} \right] \right\}}{\prod_{y=k}^{n} \beta_{iy} \prod_{x=k}^{n} \prod_{y=k}^{n} \beta_{xy} \prod_{y=j}^{n} \beta_{yj}} \qquad (17)$$

For the element ψ_{ii}, $i > k$, ψ_{ij} becomes

$$\psi_{ii} = \frac{1}{\alpha'_{k+1,k+1}} \frac{\left\{ \left[\sum_{i=0}^{j} a_{ii(n-k-1)} a_{kk(n-k-j+1)} s^{n-k(n-k-j)} \{1\} \right] - \left[\sum_{i=0}^{j} a_{ik(n-k-1)} a_{kj(n-k-j+1)} s^{n-k(n-k-j)} \left\{ \begin{matrix} s^2 \\ s \\ 1 \end{matrix} \right\} \right] \right\}}{\prod_{x=k}^{n} \prod_{y=k}^{n} \beta_{xy} \Big/ \beta_{ii} \prod_{x=i+1}^{n} \beta_{xi} \prod_{y=i+1}^{n} \beta_{iy}} \qquad (18)$$

For the element ψ_{ij}, $i > k$, $\psi > k$, ψ_{ij} becomes

$$\psi_{ij} = \frac{1}{\alpha'_{k+1,k+1}} \frac{\left\{ \left[\sum_{i=0}^{j} a_{ij(n-k-1)} a_{kk(n-k-j+1)} s^{n-k(n-k-j)} \{1\} \right] - \left[\sum_{i=0}^{j} a_{ik(n-k-1)} a_{kj(n-k-j+1)} s^{n-k(n-k-j)} \left\{ \begin{matrix} 1 \\ s \end{matrix} \right\} \right] \right\}}{\prod_{x=k}^{n} \prod_{y=k}^{n} \beta_{xy} \prod_{x=i}^{n} \beta_{xi} \prod_{y=j}^{n} \beta_{yj}} \qquad (19)$$

It should be noted that the operation at the last pivotal operation about ψ_{11} the denominator will be different. At the end of the last pivotal operation, each term will be divided by ψ_{11}.

A simple proof of the steps presented in (15)-(19) is given below. A more rigorous proof is not given here for simplicity. In each case this follows by induction. For example, for rule A we commence by setting (similarly to (15))

$$\phi(k) \equiv \psi_{ij} = \frac{\left\{\frac{1}{\alpha'_{k+1,k+1}}\alpha_{ij}\alpha_{kk}\beta_{ik}\beta_{kj} - \alpha_{ik}\alpha_{kj}\beta_{ij}\beta_{kk}\right\}}{\beta_{ij}\prod_{x=k}^{n}\prod_{y=k}^{n}\beta_{xy}\prod_{y=k}^{n}\beta_{iy}\prod_{x=k}^{n}\beta_{xj}} \qquad (20)$$

where $i < k$, $j < k$.

Suppose $\phi(m)$ is true, then when $k = m-1$ (since the operations proceed in sequence from row n back to row 1),

$$\psi_{ij} = \frac{\beta_{mm}}{\alpha_{mm}}\frac{\alpha_{ij}}{\beta_{ij}} \cdot \frac{\alpha_{m-1,m-1}}{\beta_{m-1,m-1}} - \frac{\alpha_{i,m-1}}{\beta_{i,m-1}} \cdot \frac{\alpha_{m-1,j}}{\beta_{m-1,j}} \qquad (21)$$

where β's denote "current" denominators. Substituting for these, elementary processes show that

$$\psi_{ij} = \phi(m-1) \qquad (22)$$

so that if ϕ is true for m, it is true for m-1. But when $k = n$,

$$\begin{aligned}\psi_{ij} &= \left\{\frac{\alpha_{ij}\alpha_{nn}}{\beta_{ij}\beta_{in}} - \frac{\alpha_{in}\alpha_{nj}}{\beta_{in}\beta_{nj}}\right\} \\ &= \frac{\alpha_{ij}\alpha_{nn}\beta_{in}\beta_{nj} - \alpha_{in}\alpha_{nj}\beta_{ij}\beta_{nn}}{\beta_{ij}\beta_{nn}\beta_{in}\beta_{nj}} \\ &= \phi(n) \qquad (23)\end{aligned}$$

so that ϕ is true for all k.

By substituting for the numerator $\alpha_{ij}\alpha_{nn}\beta_{in}\beta_{nj} - \alpha_{in}\alpha_{nj}\beta_{ij}\beta_{nn}$ in terms of previous numerators, it is easily shown that this is divisible by $\alpha'_{k+1,k+1}$. Similar proofs apply for rules B to E. The previously mentioned equations (15)-(19) could be applied to active networks by simply changing the second term in (C) and (D) to a negative value. This approach should, however, take advantage of symmetry.

III. NUMERICAL EXAMPLE

Consider the circuit of Fig. 1. The transfer function obtained using the method described in Section II with single precision arithmetic was

$$T(s) = \frac{1.0001s^4 + 1.2499999s^2 + 0.24999}{1.0001s^4 + 0.4997s^3 + 1.75001s^2 + 0.5001s + 0.249989}$$

compared with the actual value of

$$T(s) = \frac{s^4 + 1.25s^2 + 0.25}{s^4 + 0.5s^3 + 1.75s^2 + 0.5 + 0.25}$$

The computing time (execution time) was 24 seconds on an IBM 360/50.

IV. CONCLUSIONS

An accurate method for the inversion of a nodal matrix symbolically has been presented. It compared well with actual results. With the added advantage of improved inverse Laplace techniques [8] the symbolic formulation of the circuit by a nodal approach might provide a very good alternative to the computer-aided analysis of linear active networks. Further work is, however, required to extend this approach to very large circuits using symmetry and sparse matrix techniques.

V. REFERENCES

1. R. B. Shipley and D. Coleman, "A new direct matrix inversion method", AIEE Trans., Part I, Vol. 78, pp. 568-572, 1959.

2. D. A. Calahan, Computer-Aided Network Design, preliminary edition, McGraw-Hill, 1968 (see pp. 32-33).

3. M. L. Liou, "A novel method of evaluating transient response", Proc. IEEE, Vol. 54, pp. 20-23, January 1966.

4. W. E. Thomson, "Evaluation of transient response", Proc. IEEE, Vol. 54, p. 1584, 1966.

5. M. L. Liou, Time and Frequency Domain Analysis of Linear Time-Invariant Systems, Chapter 4, Systems Analysis by Digital Computer, eds. F. F. Kuo and J. F. Kaiser, John Wiley and Co., 1966.

6. K. A. Muraleedharan and K. A. Stromsmoe, "On computation of transient response", Proc. IEEE, Vol. 60, No. 8, p. 995, August 1972.

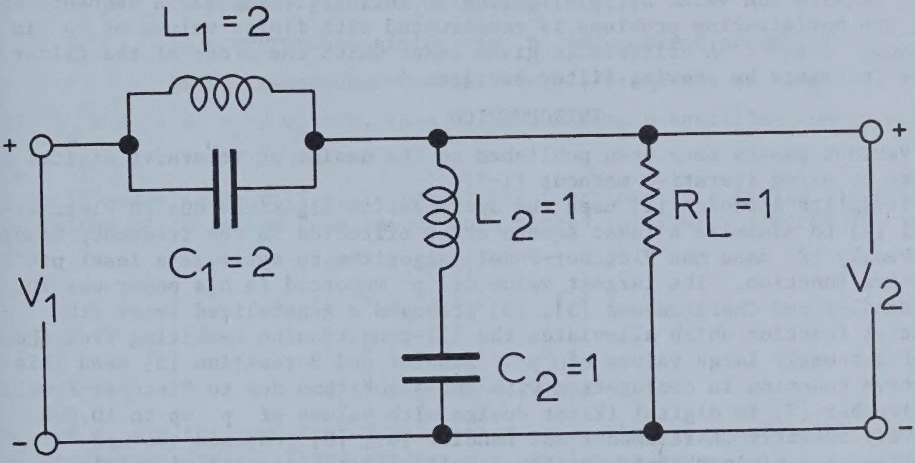
7. M. Siverberg, "A new method of solving state variable equations permitting large step sizes", Proc. IEEE (letters), pp. 1352-1353, August 1968.

8. N. B. Rabbat, "Computer-aided transient response of multiple coupled unequal transmission lines", IEEE International Symposium on Circuit Theory, Toronto, pp. 217-220, April 1973.

9. G. E. Alderson and P. M. Liu, "Computer generation of symbolic network functions -- a new theory and implementation", IEEE Trans. on Circuit Theory, Vol. CT-10, No. 1, January 1973.

10. T. F. Gatts, "TAPLAN - a topological analysis program for linear active networks", Manual, University of Alabama, University, Alabama, April 1973.

11. T. F. Gatts, private communication, June 24, 1973.


$L_1=2$
$C_1=2$
$L_2=1$
$C_2=1$
$R_L=1$
V_1
V_2
+
-
+
-

Units: ohms, henrys, & farads

Fig. 1

MINIMAX OPTIMIZATION OF RECURSIVE DIGITAL FILTERS

CHRISTAKIS CHARALAMBOUS
Department of Combinatorics and Optimization
University of Waterloo
Waterloo, Ontario, Canada

Abstract. The application of two new algorithms for minimax optimization due to Charalambous and Bandler to the problem of finding the coefficients of a recursive digital filter to meet arbitrary response specifications of the magnitude or the group delay characteristics, is investigated. Unlike the original minimax algorithm due to Bandler and Charalambous in which a sequence of least pth optimization as $p \to \infty$ is taken the two new algorithms do not require the value of p to tend to infinity. Instead a sequence of least pth optimization problems is constructed with finite values of p in the range $1 < p < \infty$. A criteria is given under which the order of the filter can be increased by growing filter sections.

INTRODUCTION

Various papers have been published on the design of recursive digital filters by using iterative methods [1-3].

Steiglitz approach [1] uses the optimization algorithm due to Fletcher-Powell [4] to minimize a least square error criterion in the frequency domain.

Deczky [2] uses the Fletcher-Powell algorithm to optimize a least pth objective function. The largest value of p reported in his paper was 10.

Bandler and Charalambous [5], [6] proposed a generalized least pth objective function which alleviates the ill-conditioning resulting from the use of extremely large values of p. Bandler and Bardakjian [3] used this objective function in conjunction with the algorithms due to Fletcher-Powell and Fletcher [7] in digital filter design with values of p up to 10,000.

Very recently Charalambous and Bandler [6], [8] proposed two new algorithms for minimax approximation. Unlike their original algorithm in which a sequence of least pth optimizations as $p \to \infty$ is taken, the two new algorithms do not require the value of p to tend to infinity. Instead a sequence of least pth optimization problems with finite values of p in the range $1 < p < \infty$ is constructed. In their paper, they prove that both algorithms will converge to the minimax optimum.

This paper describes the application of the two new minimax algorithms for choosing the coefficients of a recursive digital filter to meet arbitrary specifications of the magnitude or the group delay characteristics. Based on the generalized least square optimum the order of the filter can be increased by growing filter sections to meet the design specifications. The pole inversion technique due to Steiglitz [1] was used to meet the stability constraints.

THEORY

The Form of the Transfer Function H(z) [1]. The transfer function of a digital filter under consideration is given by

$$H(z) = A \prod_{k=1}^{K} \frac{1+a_k z^{-1}+b_k z^{-2}}{1+c_k z^{-1}+d_k z^{-2}} \tag{1}$$

For stability of the filter, the poles of the transfer function should lie within the unit circle in the z-domain.

This work was supported by the National Research Council of Canada, through a Postdoctorate Fellowship to the author, under grant C154 and through the Department of Combinatorics and Optimization of the University of Waterloo.

The Error Functions [5]. Define real error functions related to the "upper" and "lower" specifications, respectively as follows

$$e_u(\underset{\sim}{\phi},\psi) \triangleq w_u(\psi)(F(\underset{\sim}{\phi},\psi) - S_u(\psi)) \tag{2}$$

$$e_\ell(\underset{\sim}{\phi},\psi) = w_\ell(\psi)(F(\underset{\sim}{\phi},\psi) - S_\ell(\psi)) \tag{3}$$

where

$F(\underset{\sim}{\phi},\psi)$ is the approximating function (actual response)

$S_u(\psi)$ is an upper specified function (desired response bound)

$S_\ell(\psi)$ is a lower specified function (desired response bound)

$w_u(\psi)$ is an upper positive weighting function

$w_\ell(\psi)$ is a lower positive weighting function

$\underset{\sim}{\phi}$ is a vector containing the n independent parameters

ψ is an independent variable (e.g., frequency or time).

If $S_u = S_\ell = S$, $w_u = w_\ell = w$, then we have a single specified function and $e_u = e_\ell = e$.

In practise we will evaluate all the functions at a finite discrete set of values of ψ taken from one or more closed intervals. Therefore, we will let

$$\begin{aligned} e_{ui}(\underset{\sim}{\phi}) &= e_u(\underset{\sim}{\phi},\psi_i) \qquad i \in I_u \\ e_{\ell i}(\underset{\sim}{\phi}) &= e_\ell(\underset{\sim}{\phi},\psi_i) \qquad i \in I_\ell \end{aligned} \tag{4}$$

where

$$\begin{aligned} I_u &\triangleq \{1, 2, \ldots, n_u\} \\ I_\ell &\triangleq \{n_u+1, n_u+2, \ldots, n_u+n_\ell\} \end{aligned} \tag{5}$$

n_u and n_ℓ are the number of sample points over the upper and lower specifications respectively. Therefore for our problem we have

$$z_i = e^{j\psi_i\pi} \tag{6}$$

where

$$\psi_i = \frac{2f_i}{f_s} \tag{7}$$

f_s is the sampling frequency, f_i is the value of the frequency at the ith point and $\underset{\sim}{\phi} = [a_1 b_1 c_1 d_1 \ldots a_K b_K c_K d_K A]^T$. Thus $n = 4K + 1$.

The Approximating Function. The approximating function in our case will be either the amplitude or the group delay of the digital filter. From (1)

$$H(\underset{\sim}{\phi}, e^{j\theta}) = A \prod_{k=1}^{K} \frac{a_k + (1 + b_k)\cos\theta + j(1 - b_k)\sin\theta}{c_k + (1 + d_k)\cos\theta + j(1 - d_k)\sin\theta} \tag{8}$$

where $\theta = \pi\psi$. After some algebraic manipulations the amplitude of the filter is given by

$$|H(\underset{\sim}{\phi},\theta)| = A \prod_{k=1}^{K} \left(\frac{1+a_k^2+b_k^2+2b_k(2\cos^2\theta-1) + 2a_k(1+b_k)\cos\theta}{1+c_k^2+d_k^2+2d_k(2\cos^2\theta-1) + 2c_k(1+d_k)\cos\theta}\right)^{\frac{1}{2}} \tag{9}$$

The group delay is given by

$$\tau(\underset{\sim}{\phi},\ \theta) = -\frac{d}{d\theta}[\angle H(\underset{\sim}{\phi},\ e^{j\theta})]$$

$$= -\sum_{k=1}^{K} \frac{1-b_k^2+a_k(1-b_k)\cos\theta}{1+a_k^2+b_k^2+2b_k(2\cos^2\theta-1)+2a_k(1+b_k)\cos\theta}$$

$$+ \sum_{k=1}^{K} \frac{1-d_k^2+c_k(1-d_k)\cos\theta}{1+c_k^2+d_k^2+2d_k(2\cos^2\theta-1)+2c_k(1+d_k)\cos\theta} \tag{10}$$

Minimax Objective. For the minimax approximation problem we want to minimize the following objective function

$$M_f(\underset{\sim}{\phi}) = \max_{i,j}[e_{ui}(\underset{\sim}{\phi}),\ -e_{\ell j}(\underset{\sim}{\phi})] \quad i\in I_u,\ j\in I_\ell \tag{11}$$

It is important to note that if $M_f(\underset{\sim}{\phi}) > 0$ then the specifications are violated and if $M_f(\underset{\sim}{\phi}) < 0$ then the specifications are satisfied. If $M_f(\underset{\sim}{\phi}) = 0$ then the specifications are just met. Let us define for convenience,

$$f_i(\underset{\sim}{\phi}) \triangleq \begin{cases} e_{ui}(\underset{\sim}{\phi}) & i \in I_u \\ -e_{\ell i}(\underset{\sim}{\phi}) & i \in I_\ell \end{cases} \tag{12}$$

$$I \triangleq I_u \cup I_\ell \tag{13}$$

then our objective function becomes $M_f(\underset{\sim}{\phi}) = \max\limits_{i\in I} f_i(\underset{\sim}{\phi})$.

The problem of minimax optimization consists of finding a point $\overset{\vee}{\underset{\sim}{\phi}}$ such that $M_f(\overset{\vee}{\underset{\sim}{\phi}}) \le M_f(\underset{\sim}{\phi})$ for all points $\underset{\sim}{\phi}$ at least in the neighbourhood of $\overset{\vee}{\underset{\sim}{\phi}}$. Very recently Charalambous and Bandler proposed two algorithms to solve this problem [6], [8].

The Algorithms.

1. Assume the starting point $\underset{\sim}{\phi}^o$ is given; set $\xi^1=\min[0,M_f(\underset{\sim}{\phi}^o)+\varepsilon],r=1$. ε is a small positive number.
2. Minimize, with respect to $\underset{\sim}{\phi}$ the function

$$U(\underset{\sim}{\phi},\xi^r) = M(\underset{\sim}{\phi},\xi^r)\left(\sum_{i\in L(\underset{\sim}{\phi})}\left(\frac{f_i(\underset{\sim}{\phi}) - \xi^r}{M(\underset{\sim}{\phi},\ \xi^r)}\right)^q\right)^{\frac{1}{q}} \quad \text{for } M(\underset{\sim}{\phi},\xi^r) \neq 0 \tag{14}$$

$$= 0 \quad \text{for } M(\underset{\sim}{\phi},\ \xi^r) = 0$$

where

$$M(\underset{\sim}{\phi},\xi^r) \triangleq \max_{i\in I}(f_i(\underset{\sim}{\phi}) - \xi^r) = M_f(\underset{\sim}{\phi}) - \xi^r \tag{15}$$

$$q = p \operatorname{sign} M(\underset{\sim}{\phi},\ \xi^r)\begin{cases} p>1 & \text{if } M(\underset{\sim}{\phi},\ \xi^r) > 0 \\ p\ge 1 & \text{if } M(\underset{\sim}{\phi},\ \xi^r) < 0 \end{cases} \tag{16}$$

and

$$L(\underset{\sim}{\phi}) = \begin{cases} J(\underset{\sim}{\phi}, \xi^r) \triangleq \{i | f_i(\underset{\sim}{\phi}) - \xi^r \geq 0 \text{ , } i \in I\} & \text{if } M(\underset{\sim}{\phi}, \xi^r) > 0 \\ I \quad \text{if } M(\underset{\sim}{\phi}, \xi^r) < 0 & \end{cases} \tag{17}$$

3. Let $\check{\underset{\sim}{\phi}}^r$ denote the optimum parameter vector at the rth step. If Algorithm 1 go to 4; otherwise, if $M(\check{\underset{\sim}{\phi}}, \xi^r)<0$ go to 4; else set

$$\xi^{r+1} = \xi^r + \lambda^r M(\check{\underset{\sim}{\phi}}^r, \xi^r) = (1 - \lambda^r)\xi^r + \lambda^r M_f(\check{\underset{\sim}{\phi}}^r) \tag{18}$$

where $0 < \lambda^r < 1$, and go to 5.

4. Set $\xi^{r+1} = M(\check{\underset{\sim}{\phi}}^r) + \varepsilon$. $\quad$ (19)

5. Convergence criterion: If $|\xi^{r+1} - \xi^r| < \eta$, stop; else set $r = r + 1$ and go to 2. η is a small prescribed positive number.

Charalambous and Bandler proved that both algorithms will converge to the minimax optimum. It is important to note that for both algorithms the value of p can be kept constant in the range $1<p<\infty$ unlike the usual case where p has to tend to infinity.

If $f_i(\underset{\sim}{\phi})$ for $i \in I$ are continuous with continuous first partial derivatives then the objective function (14) is continuous with continuous first partial derivatives (except when $M(\underset{\sim}{\phi}, \xi^r) = 0$ and two or more maxima are equal in which case U is continuous but the gradients are discontinuous) and therefore very efficient gradient methods can be used to optimize (14) such as the one proposed by Fletcher [7].

In algorithm 1 we try to push the maximum away from the level ${}_{\vee}\xi^r$ at the rth step and in algorithm 2 we try to predict the value of $M_f(\check{\underset{\sim}{\phi}})$ by increasing the value of ξ^r from zero appropriately. The reason for the inclusion of ε is to avoid having $M = 0$.

STABILITY

The poles of the transfer function $H(\underset{\sim}{\phi}, z)$ must lie inside the unit circle in the z-domain, in order for the filter to be stable. Let us suppose that $H(\underset{\sim}{\phi}, z)$ has a real pole at $z = \alpha$. Consider the function [1]

$$P(z) = \frac{z-\alpha}{z-\frac{1}{\alpha}} \tag{20}$$

then

$$|P(z)| = |\alpha| \quad \text{when} \quad |z| = 1 \text{ .}$$

If we have complex poles these must occur in pairs as complex conjugates and the function

$$P(z) = \left(\frac{z-z_p}{z-\frac{1}{z_p^*}}\right)\left(\frac{z-z_p^*}{z-\frac{1}{z_p}}\right) \tag{21}$$

has the following property: $|P(z)| = |z_p|^2$ when $|z| = 1$. Superscript * is used to denote complex conjugate.

Equations (20) and (21) show that the magnitude characterstic of the filter remains unchanged by the pole inversion technique, as long as the value of A is replaced by the value

$$\frac{\text{Value of } A \text{ before pole inversion}}{\Pi|\text{poles outside the unit circle}|} \tag{22}$$

EXAMPLES

For all the examples considered, we used the value of $p=2$ for $r=1$, $p=10$ for $r\geq 2$ and for algorithm 2, $\lambda=0.55$ for $r=2$, $\lambda=0.6$ for $r\geq 3$. The value of ε used is 10^{-8}. Stability constraints are checked before we increase the value of r and before termination. The algorithm due to Fletcher [7] was used in all examples.

Example 1: A Wide-Band Differentiator [1]. Let us suppose that we have the following design specification for the magnitute characteristic

$$S(\psi) = \psi \ , \ 0 \leq \psi \leq 1$$

$w(\psi) = 1$ and we want to find the minimax optimum for one-section design.

TABLE I

q	$\xi^r \times 10^3$	Solution $\overset{v}{\underset{\sim}{\phi}}{}^r$	Values of ψ at which maxima occur	Absolute Maxima $\times 10^3$	$M^r_f \times 10^3$	Number of Function Evaluations
2	0	-0.33494	0.	3.1570×10^{-5}	11.381	18
		-0.66506	0.25	3.0477		
		0.85374	0.6	3.8683		
		0.10189	0.85	4.9124		
		0.36622	0.95	7.4772		
			1.	11.381		
-10	11.381	-0.25305	0.	4.1955	6.8702	138
		-0.72341	0.25	6.1265		
		0.93126	0.65	6.3264		
		0.12687	0.89	6.2614		
		0.36683	0.97	6.8702		
			1.	6.7455		
-10	6.8702	-0.28075	0.	6.4091	6.5712	79
		-0.75607	0.25	6.5449		
		0.93963	0.65	6.5370		
		0.12988	0.89	6.5519		
		0.36022	0.97	6.5712		
			1.	6.5622		
-10	6.5712	-0.28092	0.	6.5501	6.5573	77
		-0.75673	0.25	6.5563		
		0.93998	0.65	6.5563		
		0.13000	0.89	6.5564		
		0.36013	0.97	6.5573		
			1.	6.5569		
-10	6.5573	-0.28093	0.	6.5565	6.5568	82
		-0.75675	0.25	6.5567		
		0.94000	0.65	6.5567		
		0.13001	0.89	6.5567		
		0.36012	0.97	6.5568		
			1.	6.5567		

TABLE II

q	$\xi^r \times 10^3$	Solution $\underset{\sim}{\overset{v}{\phi}}{}^r$	Values of ψ at which maxima occur	Absolute Maxima $\times 10^3$	$M^r_f \times 10^3$	Number of Function Evaluations
2	0	-0.33494	0.	3.1570×10^{-5}	11.381	18
		-0.66506	0.25	3.0477		
		0.85374	0.6	3.8683		
		0.10189	0.85	4.9124		
		0.36622	0.95	7.4772		
			1.	11.381		
10	6.2598	-0.23879	0.	6.4448	6.5732	90
		-0.72496	0.25	6.5438		
		0.93961	0.65	6.5351		
		0.12987	0.89	6.5501		
		0.36786	0.97	6.5732		
			1.	6.5621		
10	6.4478	-0.23833	0.	6.5158	6.5622	29
		-0.72502	0.25	6.5525		
		0.93988	0.65	6.5510		
		0.12997	0.89	6.5537		
		0.36790	0.97	6.5622		
			1.	6.5582		
10	6.5165	-0.23817	0.	6.5417	6.5586	23
		-0.72502	0.25	6.5553		
		0.93996	0.65	6.5551		
		0.13000	0.89	6.5554		
		0.36791	0.97	6.5586		
			1.	6.5572		
10	6.5418	-0.23813	0.	6.5506	6.5574	21
		-0.72502	0.25	6.5561		
		0.93998	0.65	6.5561		
		0.13000	0.89	6.5562		
		0.36792	0.97	6.5571		
			1.	6.5574		

Solution: It was decided to use 25 sample points of ψ, namely,

ψ = 0.0, 0.85 (0.05) ; i = 1,...,18

ψ = 0.87, .97 (0.02) ; i = 19,...,24

ψ = 1. ; i = 25

Starting point: $\underset{\sim}{\phi}^0 = [-0.33\ -0.67\ 0.86\ 0.102\ 0.366]^T$

Tables I and II show the progress of algorithm 1 and algorithm 2 respectively. From table I it can be seen that for $r \geq 2$, $q = -10$. Also it is interesting to note how M^r_f decreases and asymptotically approaches the minimax optimum value. From table II it can be seen that for all values of

r the value of q is positive; which means that we stayed with algorithm 2. Also it is interesting to note how M_f^r decreases and how ξ^r increases and both of them tend to the same limit.

From tables I and II it can be seen that, although after 5 optimizations the value of M_f for both algorithms agrees up to 3 significant figures (6.56×10^{-3}) the optimum parameter vectors are different. This means that we might have two minimax optima. After perturbing the optimum parameter vectors, we went back to the same points and therefore, we concluded that we have two minimax optima.

At the optimum points reached, the poles and zeros have the following values:

Algorithm 1:	poles:	-0.16852 , -0.77147
	zeros:	1.0217 , -0.74072
Algorithm 2:	poles:	-0.16852 , -0.77146
	zeros:	0.97883 , -0.74070

If no excess phase is required, the optimum reached by algorithm 2 should be used.

Figure 1 shows the optimum error curve, after 5 sequence of optimizations It is to be noted that 101 equidistant values of ψ were used for the error evaluation and plotting.

Example 2: Low-Pass Filter. Design a low pass recursive digital filter of the cascade form, for 10KHz sampling rate, with the smallest number of sections, such that is magnitude characteristic will satisfy the following specifications [3], [9].

$f = 0,900\ (100)\ ;\quad S_u(f) = 1.1\ ;\quad S_\ell(f) = 0.9$

$f = 1200\ ;\quad S_u(f) = 0.1$

$f = 1500,\ 5000\ (500)\ ;\quad S_u(f) = 0.1\ .$

Solution: The specifications can be prescribed as:

$\psi = 0.,\ 0.18\ (0.02)\ ;\ S_u(\psi_i) = 1.1,\ S_\ell(\psi_i) = 0.9\ ;\ i = 1,\dots,10$

$\psi = 0.24\ ;\ S_u(\psi_i) = 0.1\ ;\ i = 11$

$\psi = 0.3,\ 1.\ (0.1)\ ;\ S_u(\psi_i) = 0.1\ ;\ i = 12,\dots,19.$

Before we start solving the above problem, it is worth mentioning a very important result about the objective function (14) [10].

If the objective function (14) corresponding to the optimum point of the first optimization ($r=1$) with a particular value of p is positive (negative), then, at the optimum point of any other value of p greater than one, the objective function (14) will be positive (negative) (even with $p=\infty$, i.e. minimax). Thus, even a least square solution will tell us whether we can satisfy the specifications in the minimax sense.

Using the above important result we can solve our problem in the following way:

Find the least squares optimum ($r=1$) with one-section. If the objective function is negative (i.e. specification satisfied) increase the value of p to 10 (not necessarily) and carry on with algorithm 1. If the objective function is positive (specification violated), reset $\xi^1=0$, grow another section and do another least square optimization. This procedure is repeated until we grow enough number of sections to satisfy the specification (assuming this is possible).

Starting from the point $\phi^0 = [0\ \ 1\ -1\ \ 0.5\ \ 0.1]^T (M_f(\phi^0)=0.5)$ and using the above ideas, the results shown in table III were obtained. From table III it can be seen that after 30 function evaluations ($q=2$) it was possible to realize that the minimax optimum will not satisfy the specification (without actually calculating minimax optimum point which is time consuming). Growing another section and resetting $\xi^1=0$ the objective function became negative after 26 more function evaluations (not shown in the table), which means that a two-section filter can satisfy the specification.

TABLE III

q	$\xi^r \times 10^2$	Solution $\underset{\sim}{\check{\phi}}{}^r$	$U(\underset{\sim}{\check{\phi}}{}^r, \xi^r) \times 10^2$	$M_f^r \times 10^2$	Number of Function Evaluations
2	0	-0.85815 1.00000 -1.50027 0.70905 0.15755	24.439	15.755	30
-2	0	-1.41237 1.00000 -1.57530 0.88292 0.61803 1.00000 -1.43047 0.56271 0.02603	-1.6925	-6.2827	187
-10	-6.2827	-1.40496 1.00000 -1.58462 0.88793 -0.19570 1.00000 -1.45306 0.58177 0.03547	0.95070	-7.3536	212

Example 3 [2]. In this example we consider the synthesis of a digital all-pass filter with prescribed group delay $S(\psi) = 15\psi$, $0.1 \le \psi \le 0.9$ and $w(\psi) = 1/15\psi$.

Solution: The problem was discretized into 33 uniformly spaced sample points of ψ, namely: $\psi = 0.1, 0.9\ (0.025)$, $i = 1,\ldots,33$. For all pass digital filters

$$H(z) = A \prod_{k=1}^{K} \frac{b_k z^2 + a_k z + 1}{z^2 + a_k z + b_k} \tag{23}$$

In this case, it can be easily shown that the group delay is given by

$$\tau = 2 \sum_{k=1}^{K} \frac{1 - b_k^2 + a_k(1 - b_k)\cos\theta}{1 + a_k^2 + b_k^2 + 2b_k(2\cos^2\theta - 1) + 2a_k(1 + b_k)\cos\theta} \tag{24}$$

The approximating function is given by: $F(\underset{\sim}{\phi},\psi) = \tau(\underset{\sim}{\phi},\psi) - \tau_0$ where τ_0 is a constant group delay which is considered as a variable and $\underset{\sim}{\phi} = [a_1\ b_1\ \ldots\ a_K\ b_K\ \tau_0]^T$. Using a four-section filter and starting from the point $[-0.55\ 0.39\ 0.33\ 0.5\ 1.05\ 0.58\ 1.52\ 0{,}67\ 1.12]^T$ algorithm 1 and algorithm 2 reached the following optimum points

Algorithm 1: $\underset{\sim}{\check{\phi}}^5 = [-0.58452\ 0.39747\ 0.31472\ 0.53533\ 1.03848\ 0.57649\ 1.53043\ 0.68982\ 1.14618]^T$

Algorithm 2: $\underset{\sim}{\check{\phi}}^3 = [-0.58445\ 0.39746\ 0.31473\ 0.53515\ 1.03849\ 0.57649\ 1.53040\ 0.68977\ 1.14636]^T$.

Figure 2 shows the optimum weighted error curve and Figure 3 shows the specified function and the approximating function at the optimum point. From Figure 2 it can be seen that there is a ripple which is not active at the minimax optimum.

CONCLUSIONS

The two new algorithms for minimax approximation were found very efficient in optimizing recursive digital filters. Based on the generalized least square optimum, we immediately know whether our filter can satisfy the design specifications in the minimax sense. In practise, as the results indicate, only one optimization with p=2, and two optimizations with p=10 will give very good minimax results.

ACKNOWLEDGEMENT

The author thanks Dr.J.W.Bandler, Dr.S.S.Haykim and J.R.Popovic for useful discussions.

REFERENCES

[1] K.Steiglitz, "Computer-aided design of recursive digital filters", IEEE Trans. Audio Electroacoustics, vol. AU-18, pp.123-129, June 1970.

[2] A.G.Deczky, "Synthesis of recursive digital filters using the minimum p-error criterion", IEEE Trans. Audio Electroacoustics, vol. AU-20, pp.257-263, October 1972.

[3] J.W.Bandler and B.L.Bardakjian, "Least pth optimization of recursive digital filters", Proc. International Symp. on Circuit Theory, Toronto, Ontario, April 9-11, 1973.

[4] R.Fletcher and M.J.D.Powell, "A rapidly convergent descent method for minimization", Computer J., Vol.6, pp.163-168, June 1963.

[5] J.W.Bandler and C.Charalambous, "Practical least pth optimization of networks", IEEE Trans. Microwave Theory Tech., Vol.MTT-20, pp.834-840, December, 1972.

[6] C.Charalambous, "Nonlinear least pth approximation and nonlinear programming with applications in the design of networks and systems",Ph.D. Dissertion, McMaster University, Hamilton, Ontario, Canada, Feb. 1973.

[7] R.Fletcher, "A new approach to variable metric algorithms", Computer J., vol.13, pp.317-322, August 1970.

[8] C.Charalambous and J.W.Bandler, "New algorithms for minimax optimization' Nonlinear Programming Symp., Washington, D.C., March 14-16, 1973.

[9] M.Suk and S.K.Mitra, "Computer-aided design of digital filters with finite word length", Proc. International Symp. on Circuit Theory, North Hollywood, California, April 18-21, 1972.

[10] C.Charalambous and J.W.Bandler, "New algorithms for network optimization' IEEE Trans. Microwave Theory Tech., accepted for publication.

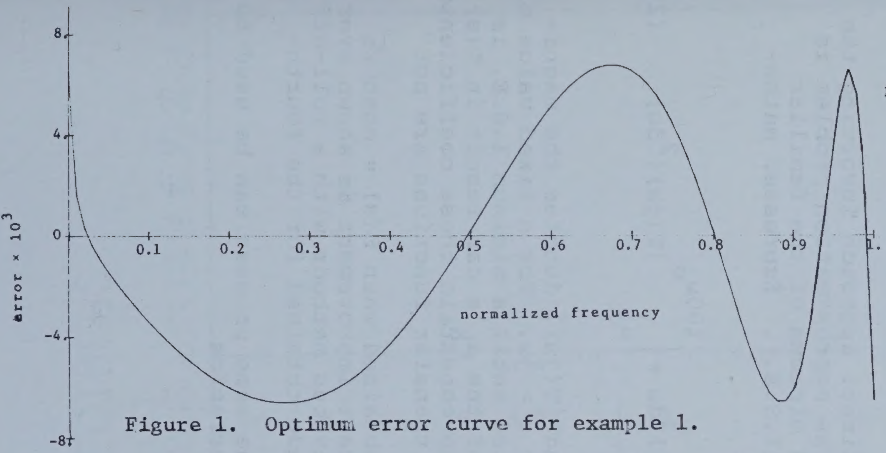


Figure 1. Optimum error curve for example 1.

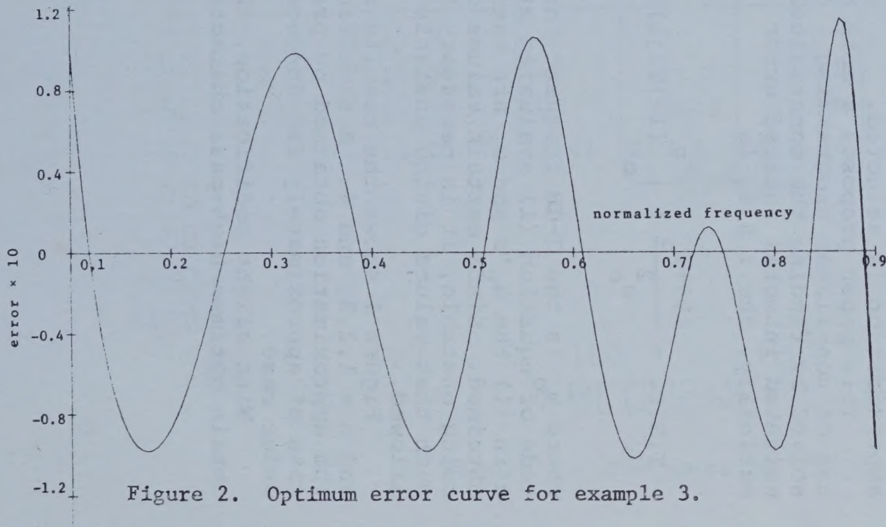


Figure 2. Optimum error curve for example 3.

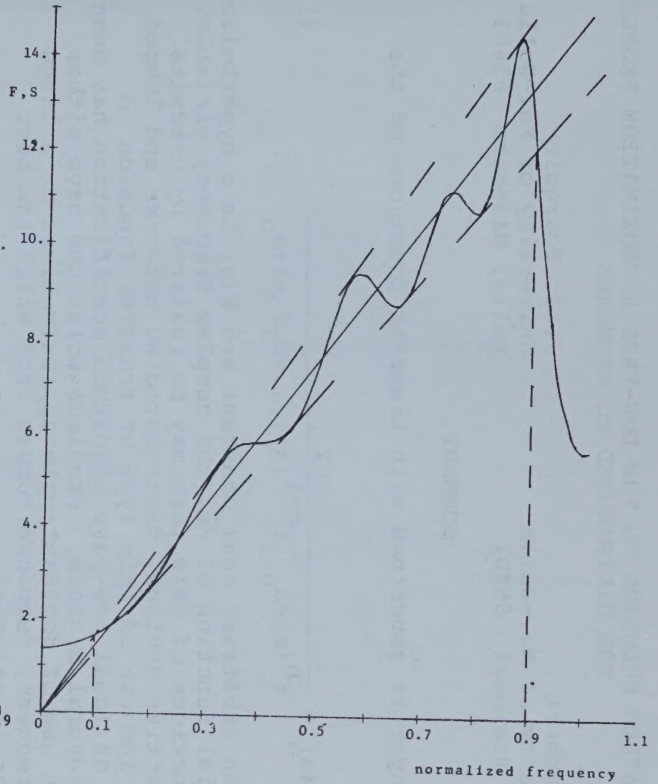


Figure 3. Optimized response for example 3.

A COMPUTER-AIDED SOLUTION TO THE LOW-PASS APPROXIMATION PROBLEM FOR DISTRIBUTED RC NETWORKS

R. I. Egbert
University of Mo.-Rolla
Rolla, Missouri 65401

J. J. Bourquin
University of Mo.-Rolla
Rolla, Missouri 65401

SUMMARY

This paper is concerned with transfer functions of the form

$$T(s) = \frac{K}{F^n(s)+a_{n-1}F^{n-1}(s)+\ldots+a_1F(s)+a_o} \tag{1}$$

where K is an arbitrary real constant and F(s) is a hyperbolic or exponential function of $\sqrt{s}$, the complex frequency variable. Transfer functions of this form may be realized by circuits containing active devices, distributed RC networks and lumped resistors. The use of this type of transfer function to approximate an ideal low-pass magnitude specification has been studied by several authors. Previous solutions have either involved the use of dominant-pole approximations or a large family of frequency response curves from which the best approximation is selected.

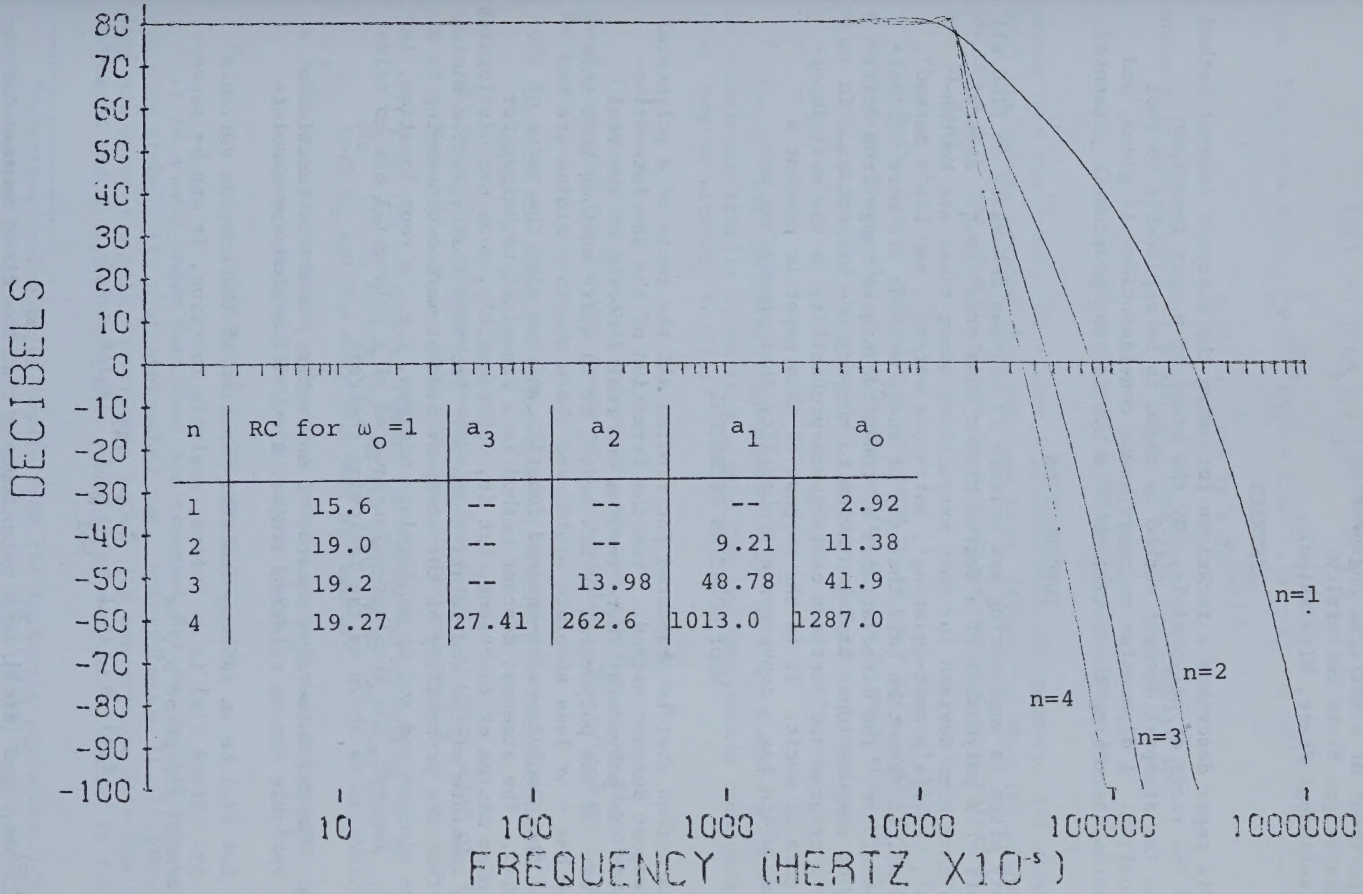
This paper proposes a more direct approach involving the use of non-linear programming. The approximation problem is solved by finding the constrained minimum of the familiar weighted integral squared error (I.S.E.). Expressed mathematically, the I.S.E. is

$$\text{I.S.E.} = \frac{100+\omega_o^2}{\omega_o^2}\left[\int_0^{\omega_o}[1-|T(j\omega)|]^2 d\omega + \int_{\omega_o}^{100\omega_o}|T(j\omega)|^2 d\omega\right] \tag{2}$$

where ω_o is the 3-dB frequency and $|T(j\omega)|$ denotes the magnitude of equation (1) evaluated at $s = j\omega$. For a fixed value of n in (1) the a_k's and ω_o are varied until a minimum I.S.E. is obtained. Since certain values of the a_k's can result in T(s) being unstable, it is necessary to constrain these coefficients such that values giving unstable transfer functions are not allowed.

Figure 1 shows the results obtained when $F(s) = \cosh\sqrt{s}$ and n = 1,2,3, and 4. A significant improvement is shown over the approximation obtained by previous methods with a roll-off rate of approximately 240 dB/decade obtained for the fourth-order case.

With slight modification, the same process can be used to obtain optimum high-pass characteristics.



n	RC for ω_o=1	a_3	a_2	a_1	a_o
1	15.6	--	--	--	2.92
2	19.0	--	--	9.21	11.38
3	19.2	--	13.98	48.78	41.9
4	19.27	27.41	262.6	1013.0	1287.0

Figure 1. Response of $|T(j\omega)|$ after optimization.

THE STEEPEST DESCENT METHOD FOR FINDING ROOTS OF POLYNOMIALS

JERREL R. MITCHELL AND WILLIE L. MCDANIEL, JR.
Department of Electrical Engineering
Mississippi State University
Mississippi State, Mississippi

ABSTRACT

This paper describes a technique for using the steepest descent method to find the roots of polynomials. By the proof of a very important theorem, the steepest descent method is shown to be applicable to root determination. A discussion on convergence considerations is given, and information describing the formation of a hybrid root program is presented.

INTRODUCTION

Very often in engineering and science it becomes necessary to find all the roots of a polynomial of a degree higher than order four. Several schemes have been devised for root extraction; among these are Newton-Raphson[1], Graeffe's root-squaring[2], Bairstows method[3], and Lin's method[4]. In general, it cannot be said that any of these methods are more suitable than the others. The disadvantages of these techniques range from difficulties of computational implementation to convergence properties. In many root program applications convergence-probability is the most important figure of merit. It is the purpose of this paper to present a technique which has a convergence probability that approaches one.

ROOT LOCATION TECHNIQUE

The method that is presented for finding all the roots of a polynomial is a steepest descent method. From the formation of the absolute-value-squared of the polynomial there results one real function of two real variables. If the polynomial is nth order, it is quite obvious that this function has n or less absolute minima and these absolute minima are the zeros of the absolute-value-squared function and are also the zeros of the polynomial. The steepest descent method is a classical technique for determining minima of functions, but it, unfortunately, does not distinguish between absolute minima and relative minima. Provided that relative minima exist, then the probability of the steepest descent method converging to a relative minimum and not to an absolute minimum, i.e., a root location, is nonzero. However, the absolute-value-squared of a polynomial has no relative minimum as is shown by the following theorem:

Theorem: The absolute-value-squared of any entire, analytic function of a complex variable has no relative minima, i.e., all minima are absolute.

Proof: Let f(s) be an entire, analytic function of the complex variable[5] $s = x + jy$. Since f(s) is an entire, analytic function, it can be represented around any point s_o, as

$$f(s) = f(s_o) + \sum_{k=1}^{\infty} \frac{1}{k!} f^{(k)}(s_o)(s - s_o)^k \tag{1}$$

where $f^{(n)}(s_o) \triangleq d^n[f(s)]/ds^n \Big|_{s=s_o}$. Suppose that the first nonzero derivative of f(s) evaluated at s_o is $f^m(s_o)$ and suppose that s is chosen sufficiently close to s_o that the series is dominated by the mth term. Then the approximate absolute value squared of f(s) is

$$F(s) \quad = \quad \left| f(s_o) + \frac{1}{m!} f^{(m)}(s_o)(s - s_o)^m \right|^2 \quad . \tag{2}$$

Let $f(s_o) = A + jB$, $\frac{1}{m!} f^{(m)}(s_o) = C\, e^{j\delta}$, $s - s_o = D\, e^{j(\theta - \delta/m)}$. Then

$$F(s) \quad = \quad \left| A + jB + Ee^{jm\theta} \right|^2 \quad . \tag{3}$$

where E is the product of D and C. Simplying (3) results in

$$F(s) \quad = \quad A^2 + B^2 + E^2 + 2E\sqrt{A^2 + B^2} \cos\,(m\theta - \phi) \tag{4}$$

where $\phi = \tan^{-1} B/A$. If either A or B or both are nonzero, (4) shows that F(s) has values greater than and less than $|f(s_o)|^2$. Thus any point at which $A \neq 0$ and $B \neq 0$ cannot be a relative maximum or a relative minimum of F(s)

If both $A = 0$ and $B = 0$ at some point $s = s_o$, F(s) must possess a minimum at this point since there exists a $\varepsilon > 0$ such that if $0 < \delta < \varepsilon$ then $|f(s_o + \delta\, \varepsilon^{j\phi})|^2 > |f(s_o)|^2$ for all allowable δ and ϕ where $0 \leq \phi \leq 2\pi$.

Since F(s) uniformly approaches f(s) as s approaches s_o and s can be chosen arbitrarily close to s_o, then the two conclusions concerning F(s) may be extended to f(s).

Q.E.D.

The preceding theorem and proof pave the way for using the steepest descent method for finding roots of a polynomial. The procedure is as follows:

(a) Form the absolute-value-squared, P(x,y), of the polynomial, P(s), where $s = x + jy$.

(b) Using the steepest descent method find a minimum point of P(x,y), e.g., (x_o, y_o). Then one root of the polynomial is $s_o = x_o + jy_o$.

(c) From the polynomial remove the factor corresponding to the root s_o. (If s_o is complex and the polynomial has real coefficients, both s_o and its conjugate may be removed).

(d) With the reduced polynomial return to step (a) and repeat the process until all roots have been found.*

PROGRAMMING HINTS AND CONCLUSION

It is a well known fact that the steepest descent method has a convergence probability that approaches one. Unfortunately it, in general, does not possess a high convergence rate. The convergence rate can be increased and a high convergence probability maintained either by forming a hybrid-root program or by using an accelerated gradient technique, e.g., the conjugate gradient technique.[6]

---

*Sometimes inaccuracies result from the factoring procedure. This can be circumvented to some extent by performing some of the last iterations in (b) on the absolute-value-squared of the original polynomial.

The general concept of the hybrid root program is to use the gradient method to find a good approximation to a root. Then this approximate value is used as the starting point in some alternate procedure for finding roots. Of course a necessary property of the alternate procedure is that its convergence probability must be large when it is provided a starting point that is a good estimate of the root, e.g., the Newton-Raphson procedure.

Even with the formation of a hybrid root program or with an accelerated gradient technique the steepest descent method has a convergence rate much below many of the other techniques. It then appears that the practical use of the steepest descent method is limited to applications where convergence probability is more important than convergence rate.

REFERENCES

1. B. James Ley, Computer Aided Analysis and Design for Electrical Engineers, Holt, Rinehart, and Winston, Atlanta, Ga., 1970, pp. 179-204.

2. David K. Cheng, Analysis of Linear Systems, Addison-Wesley Publishing Co., Reading, Mass., 1959, pp. 383-397.

3. Carl-Erik Froberg, Introduction to Numerical Analysis, Addison-Wesley Publishing Co., Reading Mass., 1969, pp. 28-32.

4. Samuel Seely, Norman H. Tarnoff, and David Holstein, Digital Computers in Engineering, Holt, Rinehart, and Winston, Inc., Atlanta, Ga., 1970, pp. 64-65.

5. Ruel V. Churchhill, Complex Variables and Applications, McGraw-Hill Book Co., New York, N.Y., 1960, p. 40.

6. R. Fletcher and C. M. Reeves, "Function Minimization by Conjugate Gradients," Computer Journal, Vol. 7, pp. 149-154, 1964.

MINIMUM FEEDBACK SYNTHESIS WITH SMALLEST NUMBER OF STATE VARIABLES FOR SOC AND MOC ASYNCHRONOUS SEQUENTIAL MACHINES

DAVIT ELHADEF
Polytechnic Institute of Brooklyn, Brooklyn, New York 11201

EDWARD J. SMITH
Polytechnic Institute of Brooklyn, Brooklyn, New York 11201

ABSTRACT

The synthesis of asynchronous sequential machines requiring a minimum number of amplifiers corresponds to minimizing the feedback index. This paper develops procedures to solve this problem for single and multiple output change machines using the smallest number of state variables. Moreover upper bounds on the number of state variables needed are derived.

A-INTRODUCTION

It is shown [2] that for every variable y_i used in the state assignment of an asynchronous sequential machine, the next state function of this variable is a function of the inputs, other state variables, and explicity of itself. The important consequence of this is the need of amplifiers in each feedback loop. Thus the problem of synthesizing circuits requiring a minimum number of amplifiers corresponds to one of minimizing the feedback index. Friedman [1] and Unger [2] have developed procedures to solve this problem for SOC and MOC machines respectively without taking care of the number of state variables used. Friedes [3] introduced the Tracey [4] conditions in order to decrease the number of state variables for SOC tables only. This paper uses Huffman's Q variable synthesis idea, tables similar to those of Paull and Unger which help keep track of all possible steps, and thus the paper attacks the problem in the reverse direction. First the state variables will be derived, and then the Q variables introduced.

B-BASIC CONCEPTS

For the sake of completeness theorems and definitions that are useful for the understanding will be explicitly declared. Meanwhile familiarity with asynchronous sequential machine theory is assumed. The pair (unstable state, stable state) is called a transition pair. In an SOC table the set of states leading to the same state under an input I is called a destination set. Flow tables where the output state may change at most once after a single change in the input state is called an SOC table (single output change); when the restrictions that only one input variable at a time may change after the machine has reached a stable state then the machine is said to operate in the normal fundamental mode. If the output may change more than once, but no more than some fixed number of times following a single change in the input state, then we have an MOC (multiple output change) table. Throughout this paper SOC tables are considered to be strongly connected and to operate in the normal fundamental mode (NFM).

Theorem [1, 2]: Given any reduced SOC or MOC flow table with S_i stable states in the i^{th} input column, any sequential circuit realizing this table in which stable flow-table states are represented by stable circuit states must have feedback index $f \geq \lceil \log_2 (\max_i (S_i)) \rceil$.

C-SOC ASYNCHRONOUS SEQUENTIAL MACHINES

The algorithm developed for the SOC case can be presented as follows:

Algorithm

Step 1: Form all destination sets for each distinct column and mark those containing two blocks as essential ones.

Step 2: For each column form all possible two block partitions and keep those such that the product of the minimal number of them results in distinct destination sets. Delete those being similar to previous ones, and assign a cost number $\{\lceil \log_2 S_i \rceil - n\}$ to that column; where n is the number of block partitions deleted and S_i the maximum number of states.

Step 3: Starting from the essential ones, define $\rho_i = (\beta_i(1), \beta_i(0))$ and $\rho_i' = (\beta_i(0), \beta_i(I))$, and form Table II (analogous to Paull and Unger's) to find the intersection of the partitions having values 1 and 0 for their block respectively.

Form an additional table (Table III) having as rows the remaining partitions and as columns the intersection partitions derived from the previous table. Assign an "I" if the column is covered by the second block and an "0" if it is covered by the first block of the partition. Add one partition at a time to Table II from among those not covering any column at all, without exceeding $\lceil \log_2 S_i \rceil$ per column and delete this one from Table IV, i. e., partition 1234, 5678 is transferred from Table III to Table II.

Proceed until no partitions without coverage exist.

Step 4: Delete partitions having exactly one column of Table III as its 1-block and select as state variables those leading to it. Scan each row of the Table III to find two or more intersections (columns) having the same value, and where their union forms one of the blocks of the related row. Select from the Table II the terms forming these columns and not contradicting the ordered pairs previously selected. Repeat until $\lceil \log_2 S_i \rceil$ for each column is reached.

Step 5: Assign value 0 to the first block and I to the second block and derive y variables.

Step 6: Derive Q variables from two block partition sets used in the derivation of y variables.

In order to give a clear understanding of the algorithm the example mentioned below will be worked out:

EXAMPLE 1:

	I_1	I_2	I_3	I_4
1	2	①	8	3
2	②	4	8	8
3	4	1	5	③
4	④	④	5	8
5	5	7	⑤	3
6	⑥	⑥	5	8
7	2	⑦	8	3
8	2	6	⑧	⑧

TABLE 1

Step 1: Destination sets for each distinct column

1278	13	1278	1357
34	24	3456	2468
5	57		
6	68		
		"e"	"e"

Step 2: Possible two-block partitions such that the product of any two in the same column leads to the destination sets for that column.

127834,56	1324,5768	1278,3456	1357,2468
12785,346	~~1357,2468~~		
12786,345	1368,2457	"e"	"e"

Cost number $\longrightarrow$ (-1)

Step 3: Table similar to Paull and Unger's in order to find the intersection of two block partitions. I denoted the partition itself and Φ the partition (12345678,); the first block of the partition has the value 0, and the second one the value 1.

	1278,3456	3456,1278	1357,2468	2468,1357	1324,5678	5678,1324
1278,3456	I					
3456,1278	Φ	I				
1357,2468	123578,46	134567,28	I			
2468,1357	124678,35	234568,17	Φ	I		
1324,5678	123478,56	123456,78	123457,68	123468,57	I	
5678,1324	125678,34	345678,12	135678,24	245678,13	Φ	I

Table II

Table having as rows the remaining two block partitions and as columns the intersection partitions derived from the previous table.

	46	28	35	17	56	78	68	57	34	12	24	13
127834,56		0		0	$\boxed{\boxed{1}}$	0			0	0	0	0
12785,346	$\boxed{1}$	0		0		0		0	$\boxed{1}$	0		
12786,345		0	1	0		0	0		1	0		

Table III

Step 4: Partition 123478,56 is deleted because one of its blocks is the same as column 56. State variables having products that forms the mentioned partition (123478,56) are: $y_1 = 1278,3456$ $y_3 = 1234,5678$

Columns 46 and 34 form the second block of (12578,346) and have the same value 1. Hence the smallest number of state variables is four and can be chosen as follows

$y_1 = 1278, 3456$

$y_2 = 1357, 2468$

$y_3 = 1234, 5678$

$y_4 = 12578, 346$

Step 5: Resulting code:

y_1	y_2	y_3	y_4
0	1	0	0
0	0	0	0
1	1	0	1
1	0	0	1
1	1	1	0
1	0	1	1
0	1	1	0
0	0	1	0

Step 6: Q variables for different inputs

$I_1 = y_4$ and $y_1 \, y_3$

$I_2 = y_2$ and y_3

$I_3 = y_1$

$I_4 = y_2$

Next state equations for Table 1 with minimum feedback index and smallest possible number of state variables

$$
\begin{aligned}
Q_1 &= I_1 \, y_4 + I_2 \, y_2 + I_3 \, y_1 + I_4 \, y_2 \\
Q_2 &= I_1 \, y_1 \, y_3 + I_2 \, y_3 \\
Y_1 &= I_1 \, (q_1 + q_2) + I_2 \, q_1' + I_3 \, q_1 + I_4 \\
Y_2 &= I_1 \, q_1' \, q_2 + I_2 \, q_1 + I_3 \, q_1 + I_4 \, q_1 \\
Y_3 &= I_1 \, q_2 + I_2 \, q_1' + I_3 \, q_2 + I_4 \, q_1 \\
Y_4 &= I_1 \, q_1 + I_2 \, q_1 + I_4 \, q_1
\end{aligned}
$$

D-MOC ASYNCHRONOUS SEQUENTIAL MACHINES

In the previous section an algorithm was developed for the case of SOC tables. Before extending this result for MOC tables it will be useful to introduce the following definitions.

Any transition pair of a flow table having the form $N(i, I) = k$. $N(k, I) = k$, $i \rightarrow k$ is called a simple transition.

Any transition set having the form $N(i, I) = k \quad N(k, I) = \ell \quad N(\ell, I) = \ell$ is called a complex transition and will be denoted by $i \rightarrow k \rightarrow \ell$. A principal transition is the complex transition having the maximum length. The length of the transition is equal to the number of different transitions contained in it. A secondary transition is any complex transition that merges to the principal one in any state different than the stable one of the former. A non-common path transition is the part of any secondary transition different from the principal one including the common state. A destination set in an MOC state table is the set of states including the stable state and all states of the complex transition leading to this stable state. A partitioning variable is an internal state variable which distinguishes one transition from another. A stable state variable is the one which distinguishes destination sets. A transient state variable partitions dichotomies resulting from the complex transition or those resulting from states leading to the same stable state with different outputs.

The algorithm for the case of MOC tables is as follows:

Step 1: Derive all generalized destination sets of the MOC table separately for each input column. The upper bound for the number of stable state variables is equal to $\lceil \log_2 D_s \rceil$ where D_s denotes the number of destination sets per column.

Step 2: Derive all complex transition sets and states leading to the same stable state with different outputs.

For the transient state variables the upper bound is equal to the sum of the length of the principal transition, the length of the non-common path transition and the number of sets of states leading to the same stable state with different outputs. The above upper bound can be easily derived from the conditions developed in [3] for USTT state assignments for MOC tables.

Step 3: Select the stable state variables (y_s's) using the method previously developed for the SOC case.

Step 4: The set of transient state variables (y_T's) is selected as follows: y_T has the same value for states being at a distance i from the stable state of the complex transition. Therefore generate the y_T's in order to cover the initial state, then the initial and the second state, and so on until the merging state is reached.

For the case of states leading to the same next state with different outputs, such as $i \rightarrow j/Z_1$ $j \rightarrow j/Z_2$ $k \rightarrow j/Z_3$ select the y_T's to cover the resulting dichotomies (i, jk) and (ij, k).

If the state table contains don't care terms do not include those states (having "d-terms" as their next states) in the partitions in order to increase the coverage properly.

Step 5: Delete the y_T terms formed by any combination of the others, provided that the order is preserved and those eliminated can be realized as unate functions of the remaining stable and transient state variables.

Step 6: Derive next state equations for q and y variables after assigning the value "1" for the first block and the value "0" for the second block of all y variables.

In order to show the improvement on the number of state variables to be used the following MOC machine example taken from [2] will be briefly solved:

Example 2:

	I_1	I_2	I_3
1	①, 00	5, 00	①, 00
2	1, 10	3, 00	②, 01
3	2, 00	③, 01	7, 00
4	3, 11	④, 10	2, 01
5	2, 01	6, 01	⑤, 11
6	5, 11	⑥, 11	7, 00
7	8, 00	4, 00	⑦, 00
8	⑧, 01	6, 10	7, 00

Step 1: Generalized destination sets:

I_1	I_2	I_3
78	23	3678
123456	74	42
	1568	5
		1

Step 2: Complex transition sets and states leading to same stable state with different outputs

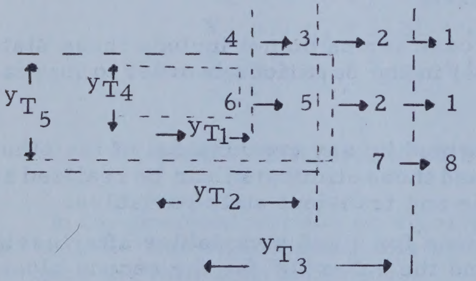
I_1	I_2
6 → 5 → 2 → 1	1 → 5 → 6
4 → 3 → 2 → 1	2 → 3
7 → 8	7 → 6
	8 → 6
	5 → 6

The upper bound of y_T's is equal to five for the input I_1 and to three for I_2 hence a total of eight y_T's is needed for this realization. However by executing the remaining steps of the algorithm it will be shown that a considerable reduction is possible in this case too.

Step 3: y_s variables derived using previous algorithm.

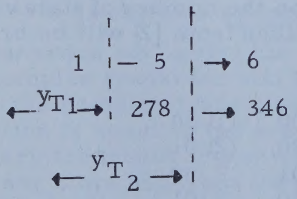
y_{1s} = 78, 123456
y_{2s} = 23, 156847
y_{3s} = 47, 123568
y_{4s} = 15, 234678
y_{5s} = 124, 35678

Step 4: Initial y_T's can be chosen as follows for input I_1:



y_{T1} = 46, 123578
y_{T2} = 3456, 1278
y_{T3} = 234567, 18
y_{T4} = 6, 1234578
y_{T5} = 56, 123478

for input I_2:



y_{T6} = 1, 2345678
y_{T7} = 15278, 346

Already the number of total y_T variables is reduced to seven.

Step 5: In order not to repeat the tables previously used the y_T variables deleted and the variables used to satisfy this step will be briefly mentioned:

$$(6,1234578) = (56,123478)(46,123578)$$

$$(2345678,1) = (234567,18) + (35678,124)$$

$$(346,12578) = (3456,1278)(46,123578) + (3456,1278)(23,145678)$$

$$(56,123478) = (15,234678)(3456,1278) + (46,123578)(35678,124)$$

Since no further simplifications is possible the total number of y_T variables used is reduced to three. Hence all y variables used have the form $y = (\beta_1(1) \, , \, \beta_0(0))$:

$$y_{s_1} = 78,123456$$

$$y_{s_2} = 47,123568$$

$$y_{s_3} = 23,145678$$

$$y_{s_4} = 15,234678$$

$$y_{s_5} = 35678,124$$

$$y_{T_1} = 46,123578$$

$$y_{T_2} = 3456,1278$$

$$y_{T_3} = 234567,18$$

Step 6: Next state equations for q's and y's can be easily derived as follows:

$$Q_1 = I_1 \, y_{s1} + I_2 \, y_{s2} + I_3 \, y_{s4}$$

$$Q_2 = I_2 \, y_{s3} + I_3 \, y_{s5}$$

$$Y_{s1} = I_1 \, q_1 + I_3 \, \overline{q}_1 \, q_2$$

$$Y_{s2} = I_2 \, q_1 + I_3 \, \overline{q}_1 \, q_2$$

$$Y_{s3} = I_2 \, q_2 + I_3 \, \overline{q}_1 \, \overline{q}_2$$

$$Y_{s4} = I_1 \, \overline{q}_1 + I_3 \, q_1$$

$$Y_{s5} = I_1 \, q_1 + I_2 \, \overline{q}_1 + I_3 \, q_2$$

$$Y_{T1} = I_2 \, \overline{q}_2$$

$$Y_{T2} = I_1 \, y_{T1} + I_2 + I_3 \, q_1 \, q_2$$

$$Y_{T3} = I_1 \, y_{T_2} + I_2 + I_3 \, (\overline{q}_1 + q_2)$$

E-CONCLUSION

Algorithms to realize minimum feedback synthesis with smalles number of state variables for SOC and MOC asynchronous sequential machines are developed and considerable improvements (specially for MOC cases) are obtained.

i. e., eight state variables used in example 2 compared to thirteen used in the previous method.

REFERENCES

1. A. D. Friedman, "Feedback in Asynchronous Sequential Circuits" IEEE Trans. on Elec. Comp., volume EC-15, Oct. 1966, pp. 740-749.

2. S. H. Unger, "Asynchronous Sequential Switching Circuits", Wiley Interscience Div. 1969.

3. A. Friedes, "State Reduction and State Assignments for Asynchronous Sequential Machines", Ph. D. Dissertation, Columbia University, 1965.

4. J. H. Tracey, "Internal State Assignments for Asynchronous Sequential Machines", IEEE-TEC, EC-15, pp. 551-560.

ACKNOWLEDGMENT

This work was supported in part by the Joint Services Contract F44620 - 69 - C - 0047.

MODULAR DATA SWITCHING NETWORKS FOR MULTIPLE COMMUNICATION DEMANDS

GERALD M. MASSON
Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, Pennsylvania 15261

ABSTRACT

This paper considers the design of data switching networks which must satisfy multiple communication demands for which it is necessary to transfer the data at one input of the network to more than one output of the network. The networks designed are composed of two-input, two-output basic switching modules with specified capabilities of connecting module inputs to module outputs. The networks are shown to be either rearrangeably nonblocking or strictly nonblocking for all classes of demands considered and an algorithm for satisfying particular demands for each network is indicated.

I. INTRODUCTION

This paper considers the design of modular data switching networks which are to provide data paths from a set of data inputs to a set of data outputs such that it is possible to transfer the data at one input to more than one output. Such data demands are called multiple communication demands and will henceforth be denoted as MCD's.

Previous work concerning the design of data switching networks has been primarily directed towards the consideration of one-to-one type, or single communication demands. This is understandable when one considers that the initial motivation for such work was the design of large-scale telephone switching networks. Beneš' classic book provides a thorough description of much of this work [1]. Recently, however, data switching networks have found important applications in non-telephone switching situations such as data reconnaissance systems, "wired city" designs, and reconfigurable computers. In these cases the network capability of being able to satisfy all single communication demands is clearly no longer sufficient. Motivated by this apparent need for a more general theory, Masson and Jordan [2-4] and Hwang [5] have extended some of the fundamental results of data switching networks. In this paper a new class of data switching networks capable of satisfying specified classes of MCD's is presented.

The networks to be presented here will be composed of two-input, two-output switching modules of two types: a type 1 basic switching module, henceforth denoted as a BSM(1), will have the capability of any one to one connection, or any subset of a one to one connection, of the module inputs to the module outputs, as shown in Figure 1(a); a type 2 basic switching module, henceforth denoted as a BSM(2), will not only have all the connecting capability of a BSM(1), but will in addition have fanout capability, as shown in Figure 1(b). Design considerations of networks composed of BSM (1)'s (where the BSM(1)'s have been called binary cells and Beta elements) for single communication demands have previously been published [6-9]. In this paper these results are initially extended to the more general MCD's, and then the use of the BSM(2)'s for these more general requirements are shown to lead to an interesting alternative network.

The networks presented in this paper will be shown to be either rearrangeably nonblocking or strictly nonblocking for the classes of MCD's considered. A data switching network is said to be rearrangeable if given any existing set of data paths through the network and any additional data path which must be provided, the existing data paths can be rearranged, if necessary, so as to provide this additional data path. A data switching network is said to be strictly nonblocking if such rearrangements are never necessary.

2. Classes of Multiple Communication Demands

MCD's can be classified according to the maximum degree of fanout required for each input to the network. We will denote the inputs to the network as the set of integers I=(1,...,N) and the outputs of the network as the set $0=(o_1,...,o_M)$. Any MCD can then be described as a mapping of 0 into I denoted by

$$f = \begin{pmatrix} o_1 & o_2 & \cdots & o_M \\ i_1 & i_2 & \cdots & i_M \end{pmatrix} \tag{1}$$

which associates with each output, o_j, the input, $f(o_j)=i_j$, $i_j \in I$, to which it is to have a data path. If a data switching network is capable of providing all the data paths required by a particular MCD, that MCD is said to be realizable by that network. In the following we will consider the realization of MCD's where each output is to be connected to some input. It is clear that if such a MCD is realizable then any subset of the connections required by that MCD is also realizable. Thus, we can rewrite (1) as

$$f = (i_1, i_2, ..., i_M) \tag{2}$$

Furthermore, in the network designs of this paper, the outputs will be connected in pairs to output subnetworks, as illustrated in Figure 2. Thus, (2) can be written as

$$\begin{aligned} f &= ((i_1, i_2), (i_3, i_4), \ldots, (i_{M-1}, i_M)) \\ &= (f_1, f_2, \ldots, f_{r_2}) \end{aligned} \tag{3}$$

where $f_j = (2_{j-1}, 2_j)$, $j=1,\ldots,r_2$, and $r_2 = \frac{M}{2}$. The case where M is not divisable by 2 will be seen to follow directly.

We will now define the two types of MCD's to be considered in this paper:

Definition: For a given I and 0, the $C(k_1,\ldots,k_N)$ class of MCD's is the set of all possible data path requirements for which there are to be at most k_1 data paths to the outputs from input 1, and there are to be at most k_2 data paths to the outputs from input 2,..., and there are to be at most k_N data paths to the outputs from input N, but to each output there is to be at most one data path from the inputs.

Example: Let $I=(1, 2, 3, 4)$ and $0=(o_1,\ldots,o_8)$. Then a MCD in the $C(2, 2, 3, 1)$ class would be

$$f = \begin{pmatrix} o_1 & o_2 & o_3 & o_4 & o_5 & o_6 & o_7 & o_8 \\ 3 & 2 & 3 & 1 & 1 & 4 & 2 & 3 \end{pmatrix} \tag{4}$$

or

$$f = ((3, 2), (3, 1), (1, 4), (2, 3)).$$

Note that this particular MCD could also be put into, say, the C(4, 4, 4, 4) class or the C(3, 3, 4, 2) class. Also, all the MCD's in the C(2, 2, 3, 1) class are certainly contained in the C(3, 3, 4, 1) class, just as all the MCD's in the C(3, 3, 4, 1) class are contained in the C(4, 4, 4, 4) class.

The other class of MCD's to be considered in this paper is the extension of this above class to the total fanout case:

Definition: For a given I and 0, the G(N, M) class of MCD's is the set of all possible data path requirements for which there can be any number of data paths to the outputs from each input, but to each output there is to be at most one data path from the inputs.

Clearly, for a given I and 0, the G(N, M) class of MCD's will contain all the MCD's of any $C(k_1,\ldots,k_N)$ class.

3. BSM(1) Networks

To design networks composed of BSM(1)'s to realize MCD's of the types described in the previous section, we will use some results from transversal theory [10] concerning sets of restricted representatives, denoted SRR's, which are discussed in the Appendix. We will consider here the case where $k_1 + \ldots k_N = M$, and results for other cases will be seen to follow with only minor modifications.

Theorem 1: For any $f=(f_1,\ldots,f_{r_2})$ corresponding to a MCD in the $C(k_1,\ldots,k_N)$ class where $k_1 + \ldots + k_N = M$, there exists a SRR involving at least $\left\lfloor \frac{k_j}{2} \right\rfloor$ and at most $\left\lceil \frac{k_j}{2} \right\rceil$ j's, j=1,...,N.[1]

Proof: We will assume that M is divisible by 2, for when it is not, we will add a hypothetical input and output for the purposes of design and then delete them in the final construction. Thus, consider any t f_i's, $(f_{i(1)},\ldots,f_{i(t)})$. This set describes the inputs to be connected to 2t outputs. Let $(i_1,\ldots,i_s)$ be this set of inputs, where $\bigcup_j^t f_{i(j)} = (i_1,\ldots,i_s)$. By the definition of an MCD in the $C(k_1,\ldots,k_N)$ class, i_1 can be connected to at most k_{i_1} outputs,..., and i_s can be connected to at most k_{i_s} outputs. Thus, $\sum_j^s k_{i(j)} \geq 2t$, or

$$\frac{k_{i_1} + \ldots + k_{i_s}}{2} \geq t \tag{5}$$

Certainly, $\left\lceil \frac{k_i}{2} \right\rceil \geq \frac{k_i}{2}$, i=1,...,N, so by (5),

$$\left\lceil \frac{k_{i_1}}{2} \right\rceil + \ldots + \left\lceil \frac{k_{i_s}}{2} \right\rceil \geq t \tag{6}$$

Next, let $x = \frac{M}{2} - \sum_i^N \frac{k_i}{2}$. Then by (5),

$$\left\lfloor \frac{k_{i_1}}{2} \right\rfloor + \ldots + \left\lfloor \frac{k_{i_s}}{2} \right\rfloor + x \geq t \tag{7}$$

Thus, by (6), (7), and Theorem A1 of the Appendix, the theorem is proven.

It follows from Theorem 1 that there exists a second SRR for f where if i appeared $\left\lfloor \frac{k_i}{2} \right\rfloor$ times in the first SRR, it appears $\left\lceil \frac{k_i}{2} \right\rceil$ times in the second SRR and vice versa, i=1,...,N.

---

[1] $\lceil a \rceil$ denotes the smallest integer greater than a, and $\lfloor a \rfloor$ denotes the greatest integer less than a, where a is any real number.

Using this result we can now construct a network of BSM(1)'s to realize any MCD in a given $C(k_1,\ldots,k_N)$ class. Consider some MCD in a $C(k_1,\ldots,k_N)$ class where $k_1 + \ldots + k_N = M$. We can assume without loss of generality that, of the integers $k_1,\ldots,k_N$, we have that $k_1,\ldots,k_p$ are each divisible by 2, and that $k_{p+1},\ldots,k_N$ are each not divisible by 2. Now, if we choose some output to be connected in this MCD to input p + 1 and some other output to be connected in this MCD to input p + 2, we can for the purposes of constructing a network consider both of these outputs to be assigned by the MCD to some hypothetical input. In other words, we can now consider a MCD in the $C(k_1,\ldots,k_p, k_{p+1} - 1, 2, k_{p+2} - 1, k_{p+3},\ldots,k_N)$ where $k_1,\ldots,k_p, k_{p+1} - 1, 2, k_{p+2} - 1$ are each divisible by 2, and $k_{p+3},\ldots,k_N$ are each not divisible by 2. Repeating the above on the remaining k_i's not divisible by 2 results in the consideration of a MCD in the $C(k_1,\ldots,k_p, k_{p+1} - 1, 2, k_{p+2} - 1,\ldots,k_{N-1} - 1, 2, k_N - 1)$ class. Then from Theorem 1 it is easily seen that the network of Figure 3 will realize the original MCD in the $C(k_1,\ldots,k_N)$ class where each middle subnetwork must be capable of realizing any MCD in the $C(\frac{k_1}{2},\ldots,\frac{k_p}{2}, \frac{k_{p+2} - 1}{2}, 1, \frac{k_{p+2} - 1}{2},\ldots,\frac{k_{N-1} - 1}{2}, 1, \frac{k_N - 1}{2})$ class. Note that the outputs are connected in pairs to BSM(1)'s except for outputs o_{M-1} and o_M which are connected directly to the middle subnetworks. This is possible because middle subnetwork 1 realizes one of the SRR's of f and middle subnetwork 2 realizes the other SRR of f. Since one of the SRR's contains the connection to o_{M-1} and the other SRR contains the connection to o_M, we can, with no loss of connecting capability to the network, designate middle subnetwork 1 to always provide the data path to o_{M-1} and middle subnetwork 2 to always provide the data path to o_M. We can now repeat the above on the two middle subnetworks, and so forth until the entire network has been constructed. We will call the network resulting from the above procedure a N_1 network. By the above, we can then state the following:

Theorem 2: Any MCD in a $C(k_1,\ldots,k_N)$ class where $k_1 + \ldots + k_N = M$ is realizable with a N_1 network.

Note that if $k_1 = k_2 = \ldots = k_N = 1$, then the above procedure yields the networks of Joel [6], Waksman and Kautz et. al. [7, 8], and Opferman and Tsao-Wu [9].

Examples: Consider the design of a N_1 network to realize all MCD's in the C(2, 2, 3, 1) class where M=8. By the above, we first design the outer stages as shown in Figure 5. Each middle subnetwork of Figure 5 must be capable of realizing any C(1, 1, 1, 1,) MCD. Thus, we next design the MS_1 and MS_2 networks as shown in Figure 6. Figure 7 shows the realization of the MCD corresponding to f = ((3, 2), (3, 1), (1, 4), (2, 3)).

Next, consider the design of a C(2, 2, 3) network where M=7. As suggested in the proof of Theorem 1, we design the outer stage as though we were considering the C(2, 2, 3, 1) class where M=8, which would result in the network of Figure 6. However, before we consider the design of the two middle subnetworks, we delete o_8 and input 4. This allows us to also delete the BSM(1) of the input subnetwork of Figure 6. Now we must design a C(1, 1, 1, 1) network where M=4 for MS_1 and a C(1, 1, 1, 1) network where M=3 for MS_2. Doing so results in the network of Figure 8.

In the above construction procedure, the number of BSM(1)'s in the outer stage is

$$\left\lfloor \frac{k_1 + \ldots + k_N}{2} \right\rfloor + \left\lceil \frac{M}{2} \right\rceil - 1 - \sum_{i=1}^{N} \left\lfloor \frac{k_i}{2} \right\rfloor \tag{8}$$

or

$$M - 1 - \sum_{i=1}^{N} \left\lfloor \frac{k_i}{2} \right\rfloor$$

since $k_1 + \ldots + k_N = M$. The total number of BSM(1)'s in outer stages of the two middle subnetworks can similarly be seen to be

$$M - 2 - \sum_{i=1}^{N} \left\lfloor \frac{\left\lfloor \frac{k_i}{2} \right\rfloor}{2} \right\rfloor . \tag{9}$$

Let

$$B_i = \left\lfloor \frac{k_i}{2} \right\rfloor + \left\lfloor \frac{\left\lfloor \frac{k_i}{2} \right\rfloor}{2} \right\rfloor + \left\lfloor \frac{\left\lfloor \frac{\left\lfloor \frac{k_i}{2} \right\rfloor}{2} \right\rfloor}{2} \right\rfloor + \ldots \tag{10}$$

Then the total number of BSM(1)'s in a N_1 network to realize all MCD's in a $C(k_1, \ldots, k_N)$ class where $k_1 + \ldots + k_N = M$ and $p = \lfloor \log_2 M \rfloor$ is

$$pM - \sum_{i=1}^{p} 2^i - \sum_{i=1}^{N} B_i$$

or $\qquad$ (11)

$$pM - 2^p + 1 - \sum_{i=1}^{N} B_i$$

The case where $k_1 + \dots + k_N \neq M$ can be handled with only minor modifications at the innermost subnetwork. Here we will have the case of either $p \leq 2$ inputs and $p' \geq 2$ outputs for the case where $K_1 + \dots + K_N < M$, or $p \geq 2$ inputs and $p \leq 2$ outputs for the case where $K_1 + \dots + K_N > M$. Upon the construction of the innermost subnetworks in this case, the appropriate network of Figure 9 should then be used. Expressions for the number of BSM(1)'s in the networks resulting from these asymmetrical cases can be developed. However, such expressions are quite involved and are tantamount to a description of the construction procedure. Thus, none will be given here.

To design a network of BSM(1)'s to realize any MCD in the G(N, M) class, we note that it is sufficient to design a network where from each input there is a possible data path to all M outputs, that is, a network capable of realizing all C(M,...,M) MCD's. Thus, we can state the following:

Corollary 2.1: Any MCD in the G(N, M) class is realizable with a N_1 network of BSM(1) switching elements which has been designed to realize any C(M,...,M) for the given I and O.

Finally, an algorithm for placing the BSM(1)'s into proper states for the realization of specific MCD's can be obtained based on Theorem 1. By this Theorem we see that it is sufficient to construct SRR's at each stage, starting with the outer stage and moving inwards, for the realization of a MCD. A number of efficient algorithms for doing this are known [10, 11]. Determining such a SRR specifies the BSM(1) states for the input and output sides of the stage being considered. For conciseness these algorithms will not be repeated here.

4. BSM(2) Networks

Certainly, BSM(2) switching modules could be used in place of the BSM(1) switching modules in the networks of the previous section. However, the fanout capability of the BSM(2)'s allows the consideration of some interesting alternatives to N_1 networks, in particular for the realization of G(N, M) MCD's. Accordingly, we will construct a strictly nonblocking network of BSM(2)'s for the realization of G(N, M) MCD's. To do this we will use the concept of a set of distinct representatives, denoted as SDR's, which are discussed in the Appendix. The structures of the networks to be developed here will be even-staged, as opposed to the primarily odd-staged networks considered previously. Figure 10 shows a network with $N = 2r_1$ inputs and $M = 2r_2$ outputs where input 1 is connected to IS_1 and input 2 is connected to IS_2; and input 3 is connected to IS_1 and

IS_2, and input 4 is connected to IS_3; input 5 is connected to IS_1, IS_2, and IS_3, and input 6 is connected to IS_4;...; input $2r_1 - 1$ is connected to IS_1, ..., IS_{r_1}, and input $2r_1$ is connected to $IS_{r_1 + 1}$. Note that each output subnetwork is connected by one line to each input subnetwork, and there are two outputs connected to each output subnetwork. If each input and output subnetwork can handle all G(i, j) MCD's where there are i inputs and j outputs to the subnetwork, then such a structure will be referred to as a N_2 network.

Theorem 3: A N_2 network is strictly nonblocking for all MCD's in the G(N, M) class.

Proof: Clearly, to connect two distinct inputs to the outputs of any output subnetwork, there must exist paths from these inputs to two distinct input subnetworks, and then paths from each of these input subnetworks to this output subnetwork. By the interconnection pattern between the output and input subnetworks, this latter condition is always satisfied. Let $(I_1, \ldots, I_{r_1 + 1})$ be a set of subsets where I_j denotes the inputs connected to IS_j, $j = 1, \ldots, r_1 + 1$. Thus,

$$
\begin{aligned}
I_1 &= (1, 3, 5, \ldots, 2r_1 - 3, 2r_1 - 1) \\
I_2 &= (2, 3, 5, \ldots, 2r_1 - 3, 2r_1 - 1) \\
I_3 &= (4, 5, \ldots, 2r_1 - 3, 2r_1 - 1) \\
I_4 &= (6, \ldots, 2r_1 - 3, 2r_1 - 1) \\
&\vdots \\
I_{r_1} &= (2r_1 - 2, 2r_1 - 1) \\
I_{r_1 + 1} &= (2r_1).
\end{aligned}
$$

By the definition of a SDR from the Appendix, to realize any MCD, it is sufficient to show that each f_i of f corresponding to this MCD is a SDR of at least two of the I_j of $(I_1, \ldots, I_{r_1 + 1})$. From the construction of $(I_1, \ldots, I_{r_1} + 1)$, this is obviously always the case, which proves the theorem.

5. Conclusion

We have seen two structures composed of basic switching modules for the realization of multiple communication demands. The approach used to demonstrate that such networks are capable of realizing MCD's can be used to design networks of still other structures. This is an important area for future research since the relationship of structure and algorithm complexity must be well understood when considering the design of such networkds for their increasing number of applications.

APPENDIX

Given a set of integers, $I = (1,\ldots,N)$, and a set of subsets of I, $f = (f_1,\ldots,f_r)$, a set of restricted representatives (SRR) of f is defined to be a set of elements, one from each f_i, where the element 1 appears at least a_1 times and at most a_1' times, and the element 2 appears at least a_2 times and at most a_2' times,..., and the element N appears at least a_N times and at most a_N' times, where a_i and a_i' are integers, $i = 1,\ldots,N$, and $0 \leq a_i \leq a_i'$.

Theorem A1 (Ford and Fulkerson): An SRR for $f = (f_1,\ldots,f_r)$ in which $i_j \in I$ appears at least a_j times and at most a_j' times, $0 \leq a_j \leq a_j'$, $j = 1,\ldots,N$, exists if and only if for every subset of f, say, $f_{i(1)},\ldots,f_{i(t)}$, $1 \leq t \leq r$, we have that

$$t \leq \min\left(r - \sum_{i}^{N} a_i + \sum_{X} a_i,\ \sum_{X} a_i'\right)$$

where $X = \bigcup_{j}^{t} f_{i(j)}$.

For the same sets, I and f, a set of distinct representatives (SDR) of f is defined to be a set of elements, one from each f_i, denoted by $(i_{j(1)},\ldots,i_{j(r)})$, where

$$i_{j(k)} \in f_{j(k)} \quad k = 1,\ldots,r$$

and

$$i_{j(k)} \neq i_{j(k')} \quad k \neq k'.$$

REFERENCES

1. Benes, V. E., "Mathematical theory of connecting networks and telephone traffic," Academic Press, New York, 1965.

2. Masson, G. M. and Jordan, B. W., "Realization of a class of multiple connection assignments with asymmetrical three-stage connection networks," Proceedings of the Fifth Annual Princeton Conference on Information Science and Systems, pp. 316-320, 1971.

3. Masson, G. M. and Jordan, B. W., "Generalized multi-stage connection networks," Networks, Vol. 2, pp. 191-221, 1972.

4. Masson, G. M., "Upper bounds on fanout in connection networks," IEEE Trans. Circuit Theory, Vol. CT-20, pp. 222-230, 1973.

5. Hwang, F. K., "Rearrangeability of multiconnection three-stage Clos networks," Networks, Vol. 2, pp. 301-306, 1972.

6. Joel, A., "On permutation switching networks," Bell Sys. Tech. J., Vol 47, pp. 813-822, 1968.

7. Waksman, A., "A permutation network," J. Ass. Computing Machinery, Vol. 15, pp. 159-163, 1968.

8. Kautz, W., Levitt, K., and Waksman, A., "Cellular interconnection arrays," IEEE Trans. Computers, Vol. C-17, pp. 433-455, 1969.

9. Opferman, D. C. and Tsao-Wu, N. T., "On a class of rearrangeable switching networks, Parts I and II," Bell Sys. Tech. J., pp. 1579-1600 and pp. 1601-1618, 1971.

10. Mirsky, L., "Transversal theory," Academic Press, New York, 1970.

11. Hall, M., "Distinct representatives of subsets," Bull. Amer. Math. Soc., Vol. 54, pp. 922-926, 1948.

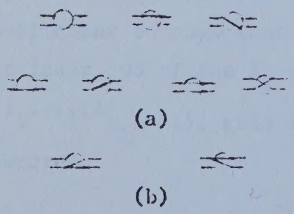


Figure 1

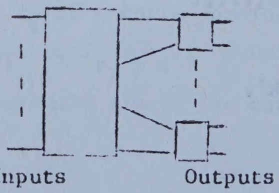


Figure 2

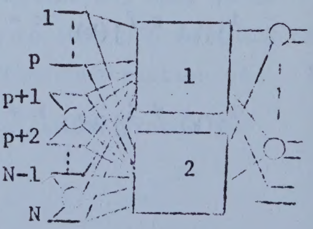


Figure 3

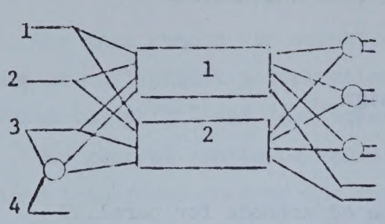


Figure 5

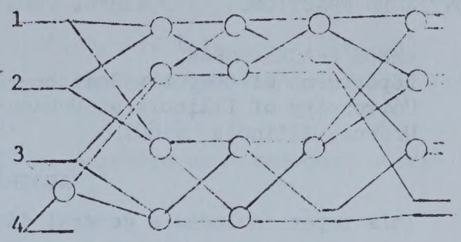


Figure 6

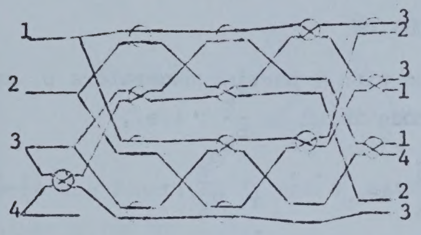

1
2
3
4
3
2
3
1
1
4
2
3

Figure 7

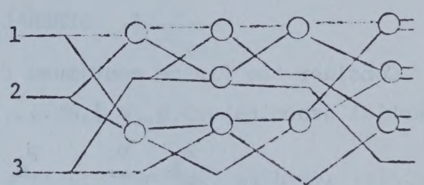
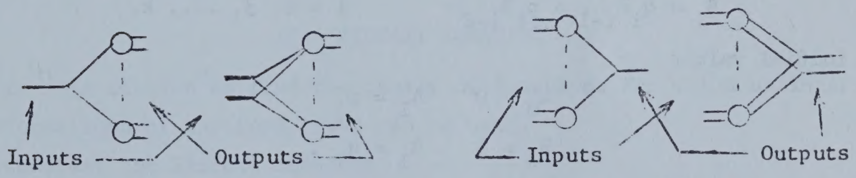


Figure 8


Inputs
Outputs
Inputs
Outputs

(a) (b)

Figure 9

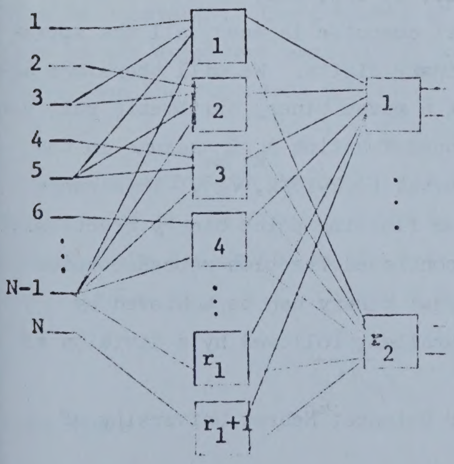

1
2
3
4
5
6
N-1
N
1
2
3
4
r1
r1+1
1
r2

Figure 10

CONTINUED FRACTIONS: A POWERFUL TOOL FOR DIGITAL ARITHMETICS

AMNON BRACHA-BARAK[†]
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

ABSTRACT

This paper includes a general discussion of methods for parallel arithmetics by using Continued Fractions. The class of functions which can be represented by a continued fraction is shown and a brief example is given to demonstrate the solution of polynomial equations.

I. GENERAL DISCUSSION

Consider the finite continued fraction with k partial numerators p_i and k partial denominators q_i = 1,2 ..., k, whose value is $\frac{A_k}{B_k}$, i.e.,

$$\frac{A_k}{B_k} = \frac{p_1}{q_1 +}\;\frac{p_2}{q_2 +}\;\frac{p_3}{q_3 +}\;\cdots\;\frac{p_k}{q_k}\;. \tag{1.1}$$

A_k and B_k are determined from the recursions:

$$\begin{aligned} A_i &= q_i A_{i-1} + p_i A_{i-2} \\ B_i &= q_i B_{i-1} + p_i B_{i-2} \qquad i = 2, 3, \ldots, k, \end{aligned} \tag{1.2}$$

with initial values:

$$\begin{aligned} A_0 &= 0 \qquad A_1 = p_1 \\ B_0 &= 1 \qquad B_1 = q_1 \;. \end{aligned} \tag{1.3}$$

It is clear that A_k and B_k can be separately and simultaneously determined in two binary arithmetic units in k-1 addition times if the p_i and q_i are chosen to be simple binary constants, i.e., 1, 1/2, etc.

Since an arithmetic unit of a digital computer is used, all the approximations are up to a finite number of binary digits. We will therefore define binary values as all the values that a given binary arithmetic unit can assume. It can be proved that the continued fraction A_k/B_k assume in the limit all the binary values over the interval $[(\sqrt{2}-1)/2, \sqrt{2}]$. This range includes [1/2, 1], the range of normalized floating point binary fractions. This property indicates that a suitable continued fraction representation exists, such that conversion to conventional binary can be achieved by repetitive use of two binary adders in parallel, followed by a division to determine the quotient $\frac{A_k}{B_k}$.

[†] Present address: Department of Computer Science, Hebrew University of Jerusalem, Jerusalem, Israel.

The main reason for selecting p_i, $q_i \in \{1/2, 1\}$, $i=1, 2, \ldots$ is that the four multiplicative operations required for each iteration in (1.2) are reduced to "shift" and "add" operations.

The general continued fraction will be regarded as a sequence of bilinear transformations of the form:

$$f_k = \frac{p_k}{q_k+f_{k+1}}, \quad k = 1, 2, \ldots, \tag{1.4}$$

where $f_k(x)$ is a function of x.

$$f_1 = \frac{p_1}{q_1} + \frac{p_2}{q_2} + \cdots + \frac{p_n}{q_n+f_{n+1}}$$

$$= \frac{A_n+f_{n+1}A_{n-1}}{B_n+f_{n+1}B_{n-1}}, \qquad n = 1, 2, \ldots,$$

where the functions A_n and B_n satisfy the recursion (1.2) with the initial values given by (1.3).

Selection rules for p and q in each iteration were studied in [3], where a complete set of such rules were developed for the quadratic equation.

II. RICCATI EQUATION

In this section we find the family of functions for which bilinear transformations of the form (1.4) can be used.

Consider the Riccati equation

$$f_1' + a_1 f_1^2 + b_1 f_1 + c_1 = 0$$

where $f_1(x)$ is a function of the variable x, and a_1, b_1 and c_1 are functions of x or constants. The property of this equation as noted by P. Wynn [4] is that if the dependent variable f_1 is replaced by the bilinear transformation (1.4) then the functions f_k, $k = 1, 2, \ldots$ also satisfy the Riccati equation

$$f_k' + a_k f_k^2 + b_k f_k + c_k = 0 . \tag{2.1}$$

We develop below the recursion for the coefficients of the $(k+1)^{st}$ equation by means of the coefficients of the k^{th} equation.

Let (2.1) be designated as the k-th Riccati equation, and assume that this is satisfied by $y_k = f_k$. From (1.4) we have

$$y_k = \frac{p_k}{q_k+y_{k+1}}, \quad p_k, q_k \in \{1/2, 1\}$$

then since

$$y'_k = \frac{-p_k y'_{k+1}}{(q_k+y_{k+1})^2}$$

we have

$$\frac{-p_k y'_{k+1}}{(q_k+y_{k+1})^2} + a_k \frac{p_k^2}{(q_k+y_{k+1})^2} + b_k \frac{p_k}{q_k+y_{k+1}} + c_k = 0 \ .$$

If we multiply by $-(q_k+y_{k+1})^2/p_k$ to normalize the coefficient of y'_{k+1}, we get

$$y'_{k+1} - \frac{c_k}{p_k} y_{k+1}^2 - (b_k + \frac{2c_k q_k}{p_k}) y_{k+1} - (a_k p_k + b_k q_k + \frac{c_k q_k^2}{p_k}) = 0 \ .$$

and the y_{k+1} will satisfy the (k+1)th Riccati equation if

$$\begin{aligned} a_{k+1} &= -\frac{c_k}{p_k} \\ b_{k+1} &= -b_k - \frac{2c_k q_k}{p_k} \\ c_{k+1} &= -a_k p_k - b_k q_k - \frac{c_k q_k^2}{p_k} \qquad k = 1,\ 2,\ \ldots \end{aligned} \tag{2.2}$$

We note that all the operations involved require no multiplication operations, since both p_k and q_k are simple binary constants.

We conclude this section by pointing to the fact that many functions have a Riccati equation and therefore can be expended as a continued fraction.

III. THE SOLUTION OF A POLYNOMIAL AS A CONTINUED FRACTION

We show now how distinct roots of a given polynomial can be approximated by using bilinear transformations.

First we develop the recursion for the coefficients of the polynomial for the $i+1^{th}$ step by using the coefficients of the i^{th} step, $i = 1, 2, \ldots$.

Let

$$P_n^1(x) = \sum_{k=0}^{n} a_k^1 x_1^k = 0 \tag{3.1}$$

be a given polynomial equation of order n. The index 1 refers to the first step in the recursion. In particular, the index 1 is not an exponent, whereas the index k is an exponent when used as a superscript.

We use a substitution of the form

$$x_i = \frac{p_i}{q_i + x_{i+1}}, \qquad i = 1, 2, \ldots, \tag{3.2}$$

where p_i, q_i are constants to be defined.

Suppose that the i^{th} step polynomial equation is of the form

$$P_n^i = a_n^i x_i^n + a_{n-1}^i x_i^{n-1} + \dots + a_1^i x_i + a_0^i = 0 , \qquad (3.3)$$

then by using substitution (3.2) we have

$$a_n^i \left(\frac{p_i}{q_i+x_{i+1}}\right)^n + a_{n-1}^i \left(\frac{p_i}{q_i+x_{i+1}}\right)^{n-1} + \dots + a_1^i \frac{p_i}{q_i+x_{i+1}} + a_0^i = 0.$$

Multiply by $(q_i + x_{i+1})^n$ to normalize the coefficient of a_n^i.

We get

$$a_n^i p_i^n + a_{n-1}^i p_i^{n-1} (q_i+x_{i+1}) + \dots + a_1^i p_i (q_i+x_{i+1})^{n-1} + a_0^i (q_i+x_{i+1})^n = 0.$$

The recursion that follows is:

$$\begin{aligned} a_n^{i+1} &= a_0^i \\ a_{n-1}^{i+1} &= a_1^i p_i + n a_0^i q_i \\ &\;\;\vdots \\ a_k^{i+1} &= \sum_{t=0}^{n-k} \binom{k+t}{t} a_{n-k-t}^i p_i^{n-k-t} q_i^t \qquad k = 0, 1, \dots, n. \end{aligned} \qquad (3.4)$$

The resulting i+1th step polynomial equation is

$$P_n^{i+1} = a_n^{i+1} x_{i+1}^n + a_{n-1}^{i+1} x_{i+1}^{n-1} + \dots + a_1^{i+1} x_{i+1} + a_0^{i+1} = 0 .$$

The method to approximate the solution of (3.1) can be used now, if an algorithm for selecting p_i and q_i $(i = 1, 2, \dots)$ is defined. This subject was extensively discussed in [2].

REFERENCES

[1] A. Bracha-Barak, "Application of Continued Fractions for Fast Evaluation of Certain Functions on a Digital Computer," IEEE Transactions on Computers (to appear).

[2] A. Bracha-Barak, "A Method for Solving Polynomial Equations by Continued Fractions," IEEE Transactions on Computers (to appear).

[3] J. E. Robertson and K. S. Trivedi, "The Status Investigations into Computer Hardware Design Based on the Use of Continued Fractions," IEEE Transactions on Computers, Vol. 22, No. 6, 1973.

[4] P. Wynn, "On Some Recent Development in the Theory and Application of Continued Fractions," SIAM Journal on Numerical Analysis, Vol. 1, pp. 177-197, 1964.

COMPUTER SCIENCE EDUCATION: A DIFFERENT MACHINE MODEL OF EFFECTIVE COMPUTATION.

MARTIN E. KALISKI, Department of Electrical Engineering,
Northeastern University,† Boston, Mass. 02115.

JACOB B. GELLER, Department of Computer Sciences,
City College of New York, New York, N.Y. 10031

This paper presents an alternative mechanical model of effective computation that the authors feel has value as an educational tool. The model is a structure with three infinite binary pushdown stacks which (i) has operators for transferring 0's or 1's from one stack to another by "pushes" and "pops" (ii) has operators for inserting 0's and 1's into one particular stack and (iii) is controlled by a program consisting of these operators and conditional branch instructions which are satisfied if the top of this same stack is a 1. Thus the model combines features of both Wang's formulation of Turing machines[1] and Minsky's program machines[2]. The model is equivalent to the above models of effective computation and has been favorably presented in a graduate course in the theory of computation. Its value lies in (i) its intuitive nature (ii) its ability to tie together traditional models of computation into a programmable form (iii) its use of pushdown stacks and (iv) its potential simulation on interactive graphical terminals.

1. Thatcher, J.W. "Self-describing Turing Machines and Self-reproducing Cellular Automata" in Essays on Cellular Automata, Arthur W. Burks, Editor, University of Illinois Press, 1970.

2. Minsky, M. Computation: Finite and Infinite Machines, Prentice-Hall, 1967.

---

† Dr. Kaliski was formerly with the Department of Computer Sciences at The City College of New York.

IMPROVING THE RELIABILITY OF FAULT-TOLERANT DIGITAL SYSTEMS BY USING MANY-VALUED LOGIC

E. G. DUCASSE
Department of Computer and Information Science
Brooklyn College, CUNY
Brooklyn, New York 11210

G. A. METZE
Department of Electrical Engineering and Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

Protective redundancy is incorporated into the design of fault-tolerant digital computers to increase their reliability and enable them to be self-repairing. In an n-tuple modular redundant system, where n is usually odd, it is common to set a threshold of n/2. It is assumed that all n modules function properly when operation begins, so that all agree initially. The system then continues working as long as n/2 or more concur, and goes down only when the threshold is not attained. This majority voting technique is essentially a binary decision. The result is accepted if n/2 or more of the modules are in accord, and rejected otherwise. If the result is accepted, it is immaterial what the extent of uniformity among the majority is, and it is therefore not recorded. However, this can have serious implications for the future operation of the system, especially in the long run. When (n+1)/2 or more of the system's modules fail, the entire system fails.

Of course it is possible to incorporate standby spare modules into the design of digital systems to replace defective units. These hybrid-redundant systems are known to be more reliable and have longer mean lives than the pure multiplexed systems. However, in addition to the advantages of these hybrid-redundant systems, there are also several problems. In order to identify a malfunctioning module, switch it out, and switch in a standby spare, two pieces of hardware in addition to the restoring organ are required. A disagreement detector is necessary to properly identify the defective module, and a switching unit must be avaliable to replace this module with a spare. Besides the expense of these additional units, there is also the difficulty of checking the disagreement detector for improper operation. This unit is particularly difficult to check, but this must be done if the system is to remain reliable.

Now suppose that one wishes to know the extent of uniformity among the agreeing majority of modules so that he is aware of how close the system is to failure, but wants to avoid the difficulties inherent in the R-S-D unit, that is, the unit composed of the restoring organ, the switching unit, and the disagreement detector. This situation can be handled very nicely by using a single multivalued gate. In an n-tuple modular redundant system, the multivalued gate would record the actual number of modules in the majority or plurality, rather than merely emitting a signal when a predetermined threshold is attained. This would give a human operator a precise knowledge of how close the system is to failure, and would allow him time to interact with the system to prevent this occurrence if he should so desire. Such an operator may wish to initiate replacement or repair procedures to prolong the system's life, or he may simply decline to act if he feels the mission's objectives can be attained in spite of the system's partial failure. Of course it is always possible for a mechanical device to assume the role of the human operator in this situation, thereby completely automating the system.

The principal advantage in using a multivalued gate to check the modules of a fault-tolerant system is that it detects trouble early. As soon as one of the units malfunctions, the output of the gate falls from its maximum value. The typical restoring organ does not reveal the failure of individual modules. If a multivalued gate is employed to reveal module failure, we are faced with a very important question: When it is determined that a unit is not functioning properly, should the less than perfect output of the system be propagated, or should a check be initiated to locate the failure? The question becomes even more difficult when one recognizes the principal causes of logic faults in a digital system. These causes are primarily either component failures in the circuits of the system or external interference with the system's operation. Prime sources of external interference are electric noise, transient variations in power supplies, and electromagnetic interference. Clearly, in the presence of one of these factors, a search for component failure would be fruitless. On the other hand, if the trouble can be traced to a defective component, initiating a search might determine the source of the error. Unfortunately, even in this situation we cannot be certain of locating the problem, for component failures are known to be of two types, permanent and intermittent, and an intermittent failure could prove very difficult to locate. Still another point to consider is that if one waits until another vote is taken, he can better determine if a fault is intermittent or permanent. Taking all these things into account and having the knowledge that he can allow the number of defective modules to approach n/2 and still have the system function properly, an operator, human or mechanical, must make the difficult decision of whether or not to interfere with the operation of the system. We do wish to point out, however, that in the absence of a multivalued logic element, the operator has no knowledge of the condition of the system and no choice of action; he can only observe helplessly until the system fails.

REFERENCES

1. Avizienis, A., Gilley, G. C., Mathur, F. P., Rennels, D. A., Rohr, J. A., and Rubin, D. K., "The STAR (Self-Testing And Repairing) Computer: An Investigation of the Theory and Practice of Fault-Tolerant Computer Design," IEEE Transactions on Computers, Vol. C-20, pp. 1312-1321, November, 1971.

2. Bouricius, W. G., Carter, W. C., Jessep, D. C. Schneider, P. R., and Wadia, A. B., "Reliability Modeling for Fault-Tolerant Computers," IEEE Transactions on Computers, Vol. C-20, pp. 1306-1311, November, 1971.

3. Hecht, H., "Figure of Merit for Fault-Tolerant Space Computers," IEEE Transactions on Computers, Vol. C-22, pp. 246-251, March, 1973.

THE DECENTRALIZED STABILIZATION AND CONTROL OF A CLASS OF UNKNOWN NON-LINEAR TIME-VARYING SYSTEMS

EDWARD J. DAVISON*
Department of Electrical Engineering
University of Toronto, Canada

ABSTRACT

The following problem is considered in this paper. Suppose a system S consists of a set of arbitrary interconnected subsystems $S_i, i=1,2,\ldots,\Omega$; is it possible to stabilize and satisfactorily control the whole system S by using only local controllers about the individual subsystems without a knowledge of the manner of the actual interconnections of the whole system? Sufficient conditions are obtained for such a result to hold true; in particular it is shown that a system S consisting of a number of subsystems S_i connected in an arbitrary way between themselves with finite gains:

$$S_i: \dot{\underline{x}}_i = \underline{A}_i(\underline{x}_i,t)\underline{x}_i + \underline{b}_i(\underline{x}_i,t)u_i,\ \underline{y}_i = \underline{c}_i'(\underline{x}_i,t)\underline{x}_i$$

where $\underline{A}_i$ and $\underline{b}_i$ have a particular structure, may be satisfactorily controlled by applying only local controllers C_i about the individual subsystems: $u_i = \underline{K}_i'(\rho)\underline{x}_i$ where $\underline{K}_i$ is a constant gain matrix with the scalar ρ appearing as a parameter, provided ρ is large enough.

I. INTRODUCTION

Systems consisting of a large number of subsystems connected in an arbitrary manner occur often in engineering, biology, ecology, etc., and a question which immediately arises in the following: is it possible to control the complete system by controlling only the subsystems of the system using only local controllers i.e. decentralized control [1], [2], the local controllers being found independent of a knowledge of the manner of the detailed interconnections of the global system? If such a class of subsystems and corresponding local controllers can be found with this property, then this would be interesting for at least two reasons:

(1) The actual design of a control system for the global system could be achieved in a very simple way; simply apply a local controller about each subsystem of the complete system.

(2) Such a control configuration would be very robust and have fail-safe

---

*Department of Electrical Engineering, University of Toronto, Toronto, Canada. This work has been supported by the National Research Council of Canada under Grant No. A4396.

properties. For example if some of the subsystems of the global system were suddenly disconnected, it would not affect the inherent global control and stability of the complete system.

It will be shown that there exists a nontrivial class of nonlinear time-varying subsystems which does possess this property. The local controller obtained in this case consists of a high gain constant state-feedback control system which may be considered as being a generalization of the well-known "high gain adaptive controller" [3], or as being a type of "model reference adaptive controller" [4]-[9]. Alternatively, the class of subsystems and local controllers taken together can be thought of as defining a "passive" system [10]. The problem studied differs from previous work on the stability of interconnected systems [11], [12] in that it is not necessary to assume a Lyapunov function exists for each subsystem and it is not necessary to assume that the gains interconnecting the subsystems are "small enough".

II. DEVELOPMENT[†]

The following theorem on stability is required to establish sufficient conditions for a solution to the problem:

Theorem 1

Given the n^{th} order system:

$$\dot{\underline{x}} = \underline{A}(\varepsilon)\underline{x} + \varepsilon\underline{f}(\underline{x},t,\varepsilon)\underline{x} + \underline{B}(\underline{x},t,\varepsilon)\underline{u} \qquad (1)$$

where $\underline{f}$ and $\underline{B}$ are continuous in $\underline{x}$ and piece-wise continuous in t and ε is a positive nonzero scalar, assume that $\underline{A}(\varepsilon)$ is stable such that $\|e^{\underline{A}(\varepsilon)t}\| < M_1 e^{-\sigma t}, \forall t \varepsilon [t_o,\infty), \forall \varepsilon\ \varepsilon(o,m)$ and that $\|\underline{f}(\underline{x},t,\varepsilon)\| < M_2$, $\|\underline{B}(\underline{x},t,\varepsilon)\| < M_3$, $\forall \underline{x} \varepsilon R^n$, $\forall t \varepsilon [t_o,\infty], \forall \varepsilon\ \varepsilon(o,m)$ where M_1, M_2, M_3, σ, m are positive scalars. Then there always exists a constant ε_o, $0<\varepsilon_o<m$ such that for any $\varepsilon<\varepsilon_o$:

(a) the system is bounded-input, bounded-output stable, and

(b) the system is globally uniformly asymptotically stable about the origin if $\underline{u} \equiv 0$.

[†] All proofs of results have been omitted in the paper; they may be found in: Davison E.J., "The Decentralized Stabilization and Control of a Class of Unknown Nonlinear Time-Varying Systems", Department of Electrical Engineering, University of Toronto, Canada, Control System Report No.7209.

Remark 1

Theorem 1 is a generalization of the well known stability results of Coddington, Levinson [13] and Bellman [14].

III. MAIN RESULTS

The following theorem gives sufficient conditions for a class of subsystems and corresponding local controllers to exist such that an unknown global system can be stabilized and be satisfactorily controlled by using only local controllers.

Consider a global system S consisting of Ω subsystems S_i of the following type:

$$S_i: \quad \begin{aligned} \dot{\underline{x}}_i &= \underline{A}_i(\underline{x}_i,t)\underline{x}_i + \underline{b}_i u_i \\ y_i &= \underline{c}_i'(\underline{x}_i,t)\underline{x}_i \end{aligned} \qquad i=1,2,\ldots\Omega \tag{2}$$

where $\underline{x}_i$ is an n_i state vector, $\underline{A}_i$ and $\underline{c}_i$ are continuous in $\underline{x}_i$ and piece-wise continuous in t, $\underline{b}_i = (0,0\ldots0,1)'$, the elements of $\underline{c}_i$ and $\underline{A}_i$ are bounded and $\underline{A}_i(\underline{x}_i,t)$ has the following structure:

$$\underline{A}_i(\underline{x}_i,t) = \begin{bmatrix} a^i_{11}(\underline{x}_i,t) & 1 & 0 & \ldots & 0 \\ a^i_{12}(\underline{x}_i,t) & a^i_{22}(\underline{x}_i,t) & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ a^i_{1n_i}(\underline{x}_i,t) & a^i_{2n_i}(\underline{x}_i,t) & a^i_{3n_i}(\underline{x}_i,t) & \ldots & a^i_{n_in_i}(\underline{x}_i,t) \end{bmatrix} \tag{3}$$

and $\underline{c}_i(\underline{x}_i,t)$ is arbitrary. Suppose these subsystems are interconnected in an arbitrary manner:

$$u_i = \sum_{\substack{j=1 \\ \neq i}}^{\Omega} g^i_j(y_j,t)y_j + u^o_i \quad , \quad i=1,2,\ldots\Omega \tag{4}$$

where g^i_j are unknown but bounded

Let $n_q \triangleq \max_{i=1,2,\ldots,\Omega} \{n_i\}$ and let $\theta_i \triangleq \begin{cases} n_i-1 \ , & n_i>1 \\ 1 \ , & n_i=1 \end{cases} \quad i=1,2,\ldots,\Omega$

The following theorem is the main result of the paper:

Theorem 2 (Main result)

Given the system S described by (2) - (4), then the local controller:

$$u^o_i = -\underline{K}_i'(\Gamma_i)\underline{x}_i + \Gamma_i^{n_i}\tilde{u}_i \quad , \quad i=1,2,\ldots,\Omega \tag{5}$$

where $\underline{K}_i$ is a constant gain matrix given by:

$$\underline{K}_i'(\Gamma_i) = \left[\Gamma_i^{n_i}, \binom{n_i}{1} \Gamma_i^{n_i-1}, \binom{n_i}{2} \Gamma_i^{n_i-2}, \ldots, \binom{n_i}{n_{i-2}} \Gamma_i^2, n_i\Gamma_1 \right] \quad (6)$$

where $\Gamma_i = \rho^{\frac{n_q-1}{\theta_i}}$ if $n_q-1 \geq 1$ and $\Gamma_i = \rho$ if $n_q-1=0$, $i=1,2,\ldots,\Omega$,

will stabilize the global system so that the controlled system is globally, uniformly, asymptotically stable about the origin if $\tilde{u}_i \equiv 0$, $i=1,2,\ldots\Omega$, and is bounded-input, bounded-output stable, independent of a knowledge of $\{g_j^i\}$ or $\{a_{k,\ell}^i\}$ provided ρ is large enough. Moreover, if $\tilde{u}_i(t)$ is piecewise continuous in t, the global controlled system will behave like:

$$\begin{aligned} \dot{\bar{\underline{x}}}_i &= \Gamma_i \tilde{\underline{A}}_i \bar{\underline{x}}_i + \Gamma_i \underline{b}_i \tilde{u}_i \\ y_i &= \underline{c}_i'(\bar{\underline{x}}_i, t)\,\mathrm{diag}(1, \Gamma_i, \Gamma_i^2, \ldots, \Gamma_i^{n_i-1})\,\bar{\underline{x}}_i \qquad i=1,2,\ldots,\Omega \end{aligned} \quad (7)$$

where $\underline{x}_i = \mathrm{diag}(1, \Gamma_i, \Gamma_i^2, \ldots, \Gamma_i^{n_i-1})\bar{\underline{x}}_i$ and

$$\tilde{\underline{A}}_i = \begin{bmatrix} 0 & 1 & 0 \ldots & 0 \\ 0 & 0 & 1 \ldots & 0 \\ \vdots & \vdots & \vdots \ldots & \vdots \\ -1, & -\binom{n_i}{1}, & -\binom{n_i}{2} \ldots & -\binom{n_i}{n_i-1} \end{bmatrix} \quad (8)$$

as $\rho \to \infty$.

Remark 2

It can be seen that as $\rho \to \infty$ the time response of the controlled global system becomes independent of the parameters $\{g_j^i\}$ and $\{a_{k,\ell}^i\}$ and becomes arbitrarily fast.

Corollary 1

Given the n^{th} order system

$$\begin{aligned} \dot{\underline{x}} &= \underline{A}(\underline{x},t)\underline{x} + \underline{b}u \\ y &= \underline{c}'(\underline{x},t)\underline{x} \end{aligned} \quad (9)$$

where $\underline{A}$, $\underline{b}$, $\underline{c}'$ have the same properties and structure of (2), then the controller

$$u = -\underline{K}'(\rho)\underline{x} + \rho^n \tilde{u} \quad (10)$$

where

$$\underline{K}'(\rho) = (\rho^n, \tbinom{n}{1}\rho^{n-1}, \tbinom{n}{2}\rho^{n-2}, \ldots, n\rho) \quad (11)$$

will stabilize the system so that the controlled system is globally, uniformly, asymptotically stable about the origin if $\tilde{u}\equiv 0$, and is bounded-input, bounded-output stable, provided Γ is large enough. If $\underline{c}'(\underline{x},t)$ is a constant matrix, and if $\tilde{u}(t)$ is of exponential class, the controlled system will behave as:

$$Y(s) = \frac{\rho^n}{(s+\rho)^n}(c_1 + c_2 s + \ldots + c_n s^{n-1})\tilde{u}(s) \tag{12}$$

as $\rho\to\infty$, where $\underline{c} = (c_1, c_2, \ldots c_n)'$. If $\underline{c}'(x,t)$ is not a constant matrix, the controlled system will behave like (7), (8) with $\Omega=1$ as $\rho\to\infty$.

An Estimate of the Gain ρ Required for the Controller

Theorem 2 states that the controller (5) will stabilize the system S provided ρ is "large enough". It is of interest to have a lower bound on the value of ρ required to ensure asymptotic stability in terms of the maximum value of the parameters $|a^i_{k,\ell}|, |g^i_j c^i_j|$ where $\underline{c}_j = \{c^i_j\}$.

Let

$$|a^i_{k,}(\underline{x}_i,t)|, |g^i_j(\underline{c}_j'\underline{x}_j,t)c^i_j(\underline{x}_i,t)| < M \quad \forall \underline{x}_i \in R^{n_i}, \; \forall t\in[t_0,\infty).$$

Theorem 3

If

$$\rho > \Omega \frac{(n_q)^{n_q+2}}{(n_q-1)^{n_q-1}} \binom{n_q-1}{\left[\frac{n_q-1}{2}\right]}^2 M \tag{13}$$

where $\left[\frac{n_q-1}{2}\right]$ = nearest integer $\geq \frac{n_q-1}{2}$, then the local controller (5) will stabilize the global system S given by (2) - (4) so that the controlled system is globally, uniformly, asymptotically stable about the origin if $u_i\equiv 0$, $i=1,2,\ldots\Omega$ and is bounded-input, bounded-output stable.

IV. SOME APPLICATIONS OF THE MAIN RESULT

Example No. 1 A Model Reference High Gain Adaptive Controller

Consider the following problem considered by Meerov [3], Dressler [5], Shackcloth [7], White [8]: A plant is described by:

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1\dot{y} + a_0 y = K_v\left(b_{n-1}\frac{d^{n-1}}{dt^{n-1}}u + \ldots + b_1\dot{u} + b_0 u\right) \tag{14}$$

where y is the output and u is the input to the system and where k_v, b_i, $i=0,1,\ldots,n-1$ are constant and known and where the coefficients

$a_i, i=0,1,\dots,n-1$ are bounded, unknown and assumed to be functions of time and $y,\dot y,\dots,y^{(n-1)}$. It is desired to find a controller for (14) so that the controlled system behaves like a given model which is as follows:

$$y_{ref}^{(m)}+\bar{a}_{m-1}y_{ref}^{(m-1)}+\dots+\bar{a}_0 y_{ref} = \bar{K}_v\left(\bar{b}_{m-1}\frac{d^{m-1}}{dt^{m-1}}u_{ref}+\dots+\bar{b}_1\dot{u}_{ref}+\bar{b}_0 u_{ref}\right)$$

where y_{ref} is the output and u_{ref} is the input to the model. In this case, the controller of Figure 1 is obtained from Corollary 1 of Theorem 2. (ρ is chosen large enough to satisfy (13)). This controller is a linear constant controller and is much simpler than the nonlinear controllers of [5], [7], [8]. (Note that this control scheme would normally be practical to implement only if the plant is minimum phase and n is small.)

Example No.2 Control of Inter-connected Systems

Consider the system S consisting of 2 subsystems S_1, S_2:

$$S_1 : \dot{\underline{x}}_1 = \begin{bmatrix} 0 & 1 \\ a_2 & a_1 \end{bmatrix}\underline{x}_1 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_1 \quad S_2 : \dot{\underline{x}}_2 = \begin{bmatrix} -2 & 1 & 0 \\ 1+\sin x_1 & 1 & 1 \\ -\cos t & -(1+\sin t) & 1 \end{bmatrix}\underline{x}_2 + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u_2 \tag{15}$$

$$y_1 = (1 \quad 0)\underline{x}_1 \qquad\qquad y_2 = (1 \quad 1 \quad 1)\underline{x}_2$$

where $a_1 = 1+\sin t+\sin x_1$, $a_2 = \cos t+\frac{1}{1+x_2^2}$, and where the two subsystems are connected in a feedback loop with $u_1 = -4y_2+u_1^0$, $u_2 = 4y_1+u_2^0$. In this case, on applying the controller of (5) from Theorem 2, the following control system is obtained:

$$\begin{aligned} u_1^0 &= -(\rho^4, 2\rho^2)\underline{x}_1 + \rho^4\tilde{u}_1 \\ u_2^0 &= -(\rho^3, 3\rho^2, 3\rho)\underline{x}_2 + \rho^3\tilde{u}_2 \end{aligned} \tag{16}$$

and on putting $\rho = 30$, the time response obtained for the controlled system with $\tilde{u}_1 \equiv 1$, $\tilde{u}_2 \equiv 0$ is almost identical to the time response of $Y_1(s) = \frac{\rho^4}{(s+\rho^2)^2}\tilde{u}_1(s)$ thereby confirming eqns. (7), (8) of Theorem 2.

V. CONCLUSIONS

This paper has considered the following problem of decentralized control: given a global system consisting of a number of subsystems which are partially unknown, is it possible to find local controllers for the subsystems such that the global system is stabilized and satisfactorily controlled without a knowledge of the detailed interconnections which occur between the subsystems? A general class of nonlinear time-varying systems which have the above property is presented.

An interesting feature that such systems possess is that of being "fail-safe" i.e. the controlled global system will still be stabilized and satisfactorily controlled, if one or more subsystems are suddenly destroyed or removed from the overall system. There are some indications that many of the fail-safe properties of biology are brought about by such feedback paths e.g. neurophysiology [15], [16].

REFERENCES

[1] Aoki M., "On Feedback Stabilizability of Decentralized Dynamic Systems", Automatica, Vol.8, No.2, March 1972, pp.163-174.

[2] Wang S.H., Davison E.J., "On the Stabilization of Multivariable Decentralized Control Systems", IEEE Trans. on Automatic Control, to appear October 1973.

[3] Meerov M.V., "Structural Bases for the Solution of Sensitivity Problems", Sensitivity Methods in Control Systems, Pergamon Press (ed. Radanovic L) 1966, pp.366-379.

[4] Nikiforuk P.N., Gupta M.M., Choe H.H., "Control of Unknown Plants in Reduced State Space", IEEE Trans. on Automatic Control, Vol.AC-14, No.5, Oct. 1969, pp.489-496.

[5] Dressler R.M., "An Approach to Model-Reference Adaptive Control Systems", IEEE Trans. on Automatic Control, Vol.AC-12, No.1, 1967, pp.75-80.

[6] Shackcloth B., "An Approach to Model-Reference Adaptive Control Systems", IEEE Trans. on Automatic Control, Vol.AC-12, No.1, 1967 pp.75-80.

[7] Shackcloth B., "Design of Model Reference Control Systems using a Lyapunov Synthesis Technique", Proc. IEE., Vol.114, No. 2, February 1967, pp.299-302.

[8] White A.J., "Analysis & Design of Model-Reference Adaptive Control Systems", Proc. IEE., Vol. 113, No. 1, January 1966, pp. 175-184.

[9] Parks P.C., "Lyapunov Redesign of Model Reference Adaptive Control Systems", IEEE Trans. on Automatic Control, July 1966, pp.362-367.

[10] Rohrer R.A., "Lumpted Network Passivity Criteria", IEEE Trans. on Circuit Theory, Vol.(T-15), No. 1, March 1968, pp. 24-30.

[11] Bailey F.N., "The Application of Lyapunov's 2nd Method to Interconnected Systems" SIAM J. on Control, Vol. 3, 1966, pp.443-462.

[12] Araki M., Kondo B., "Stability and Transient Behaviour of Composite Nonlinear Systems", IEEE Trans on Automatic Control, Aug. 1972, pp.537-541.

[13] Coddington E.A., Levinson N., Theory of Ordinary Differential Equations, 1955, pp.58-59.

[14] Bellman R., Stability Theory of Differential Equations, McGraw Hill,

1953, chapter 4.

[15] Stark L., Neurological Control Systems Studies in Bioengineering, Plenum 1968.

[16] Palmay F.v., Davison E.J., Duffin J., "The Simulation of Multi-Neurone Networks: Modelling of the Lateral Inhibition of the Eye and the Generation of Respiratory Rhythm", Dept. of Electrical Engineering, University of Toronto, Control System Report No.7301, Jan. 1973, to appear in Bulletin of Mathematical Biology.

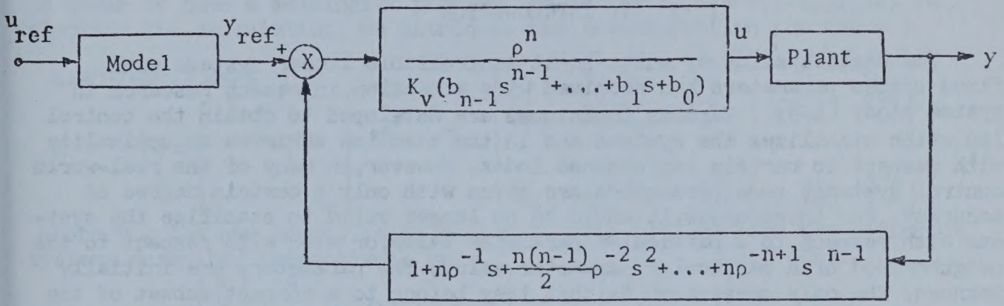

u_{ref}
Model
y_{ref}
+
-
X
$\frac{\rho^n}{K_v(b_{n-1}s^{n-1}+\ldots+b_1s+b_0)}$
u
Plant
y
$1+n\rho^{-1}s+\frac{n(n-1)}{2}\rho^{-2}s^2+\ldots+n\rho^{-n+1}s^{n-1}$

Figure 1 High Gain Model Reference Adaptive Control System

STABILIZATION OF MULTI-INPUT CONTROLLABLE LINEAR SYSTEM WITH UNKNOWN SYSTEM PARAMETERS

CHEN-YAO HO
United Technology Laboratory, Inc., Garland, Texas

ABSTRACT

Existence of a control law which can stabilize a class of multi-input controllable linear systems with unknown system parameters has been proved. The system parameters are required to lie in a compact subset of the parameter space.

I. INTRODUCTION

The stabilization of multi-input controllable linear systems with fixed system parameters has received wide attention in recent research in system study [1-3] . Various techniques are developed to obtain the control law which stabilizes the systems and in the meantime achieves an optimality with respect to certain performance index. However,in many of the real-world control systems, some parameters are given with only a certain degree of accuracy. For these cases,it would be no longer valid to stabilize the systems with respect to a particular parameter value,or even with respect to the neighborhood of a particular parameter value. The parameters are initially unknown. The only constraint is that they belong to a compact subset of the parameter space. This type of problems has been studied in [4-5] , in which the dominant type systems of scalar case are considered. The results are limited and depend on the determination of the stability region. We start with the systems with parameters assuming two different values. Then the results are generalized to the case where the system parameters assume values in a compact subset of the parameter space.

II. STABILIZATION OF MULTI-INPUT CONTROLLABLE LINEAR SYSTEMS WITH PARAMETERS ASSUMING TWO DIFFERENT VALUES

The systems under consideration are described by

$$dx/dt = Ax + Bu \tag{1}$$

where the state x is an n vector, the control u is an m vector, and A,B are constant matrices of compatiable dimensions

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{bmatrix}, \quad A_{ii} = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -a_{ii}^{1} & -a_{ii}^{2} & \cdots & -a_{ii}^{n_i} \end{bmatrix}, \quad A_{ij} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ -a_{ij}^{1} & -a_{ij}^{2} & \cdots & -a_{ij}^{n_j} \end{bmatrix} \quad \begin{matrix} i,j = 1,\cdots,m \\ i \neq j \end{matrix}$$

and

$$B = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ b_1 & & & \\ 0 & 0 & \cdots & 0 \\ \vdots & b_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & b_m \end{bmatrix} \begin{matrix} \}\, n_1 \\ \\ \}\, n_2 \\ \vdots \\ \}\, n_m \\ \end{matrix} \qquad \sum_{i=1}^{m} n_i = n \tag{2}$$

we start with the system (1) of the form (2) for two reasons. First, it could be the structure of the interconnection of m linear subsystems through inputs and outputs. Each subsystems is in a companion form. Second, as was shown in [6,7] the class of multi-input controllable linear systems having the same Brunovsky constants can be transformed into (1) of the form (2) by a nonsingular transformation and a modified input.

The elements of A, B the parameters of the systems, are assumed to belong to a compact subset C_p of $R^{n\times m}\times R^m$, where

$$C_p = \left\{ \begin{array}{l} a_{ij}^k ,\ i,j=1,\cdots,m,\ k=1,\cdots,n_j ,\ b_i ,\ i=1,\cdots,m \\ a_{ij}^k \in A_{ij}^k ,\ A_{ij}^k \text{ being a compact subset of } R ,\ i,j=1,\cdots,m,\ k=1,\cdots,n_j \\ b_{im} \leq b_i \leq b_{iM} ,\ i=1,\cdots,m \end{array} \right\}$$

where b_{im}, b_{iM} are known bounds on the parameters b_i, $i=1,\dots,m$.

We assume that the parameters of A, B take two possible values Ω_1 and Ω_2 where $\Omega_1 = \{\hat{a}_{ij}^k, i,j=1,\cdots,m, k=1,\cdots,n_j, \hat{b}_i, i=1,\cdots,m\}$ and $\Omega_2 = \{\hat{\hat{a}}_{ij}^k, i,j=1,\cdots,m, k=1,\cdots,n_j, \hat{\hat{b}}_i, i=1,\cdots,m\}$.

Mathematically, systems with parameters assuming two possible values are equivalent to two different systems with parameters Ω_1 and Ω_2 respectively. In order to have a meaningful problem, we assume $\hat{b}_i, \hat{\hat{b}}_i \neq 0, i=1,\dots,m$. To simplify the calculation, we introduce the transformation for input u

$u=H^{-1}(v-Tx)$, with

$$H = \begin{bmatrix} \hat{b}_1 & 0 & \cdots & 0 \\ 0 & \hat{b}_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \hat{b}_m \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} \hat{a}_{11}^1 \cdots \hat{a}_{11}^{n_1} & \hat{a}_{12}^1 \cdots \hat{a}_{12}^{n_2} & \cdots & \hat{a}_{1m}^1 \cdots \hat{a}_{1m}^{n_m} \\ \hat{a}_{21}^1 \cdots \hat{a}_{21}^{n_1} & \hat{a}_{22}^1 \cdots \hat{a}_{22}^{n_2} & \cdots & \hat{a}_{2m}^1 \cdots \hat{a}_{2m}^{n_m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{a}_{m1}^1 \cdots \hat{a}_{m1}^{n_1} & \hat{a}_{m2}^1 \cdots \hat{a}_{m2}^{n_2} & \cdots & \hat{a}_{mm}^1 \cdots \hat{a}_{mm}^{n_m} \end{bmatrix} \tag{3}$$

where v , the new input, is an m vector. With (3), we have two different systems described respectively by

$$\frac{dx}{dt} = \begin{bmatrix} \begin{matrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 0 \end{matrix} & & 0 \cdots 0 \\ \vdots & \ddots & \\ & & \begin{matrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & & 0 \end{matrix} \\ 0 & & \end{bmatrix} x + \begin{bmatrix} \begin{matrix} 0 \\ 0 \\ \vdots \\ 1 \end{matrix} & & 0 \\ & \ddots & \\ 0 & & \begin{matrix} 0 \\ 0 \\ \vdots \\ 1 \end{matrix} \end{bmatrix} v \tag{4}$$

and

$$\frac{dx}{dt} = \begin{bmatrix} \begin{matrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ \frac{\hat{\hat{b}}_1}{\hat{b}_1}\hat{a}_{11}^1 - \hat{\hat{a}}_{11}^1 & \cdots & & & \frac{\hat{\hat{b}}_1}{\hat{b}_1}\hat{a}_{11}^{n_1} - \hat{\hat{a}}_{11}^{n_1} \end{matrix} & \cdots & \vdots \\ \vdots & \ddots & 0 \\ & & \vdots \\ \cdots & \cdots & \frac{\hat{\hat{b}}_m}{\hat{b}_m}\hat{a}_{mm}^{n_m} - \hat{\hat{a}}_{mm}^{n_m} \end{bmatrix} x + \begin{bmatrix} 0 & & 0 \cdots 0 \\ \vdots & & \vdots \\ \hat{\hat{b}}_1/\hat{b}_1 & \ddots & 0 \\ \vdots & & \vdots \\ \cdots & & \hat{\hat{b}}_m/\hat{b}_m \end{bmatrix} v \tag{5}$$

To design a single control law v to stabilize both (4) and (5), we consider the control law v of the form

$$V = \hat{T}x = \begin{bmatrix} -C_0^{n_1}\lambda^{n_1} & -C_1^{n_1}\lambda^{n_1-1} & \cdots & -C_{n_1-1}^{n_1}\lambda & 0 & \cdots & 0 & 0 \\ 0 & & & & & \ddots & \vdots & \vdots \\ \vdots & & & & & & & \\ 0 & \cdots & & & 0 & \cdots & -C_0^{n_m}\lambda^{m} & \cdots - C_{n_m-1}^{n_m}\lambda \end{bmatrix} x \tag{6}$$

for some $\lambda>0$. With (6),(4) has a characteristic equation $\hat{G}(s) = (s+\lambda)^n$ which is a Hurwitz polynomial for any $\lambda>0$. In order to show that (6) can stabilize (5) for some $\lambda>0$, we need following Assertion[8,9]:

Assertion: $\prod_{i=1}^{m}\left(s^{n_i} + \frac{\hat{\hat{b}}_i}{\hat{b}_i} \sum_{j=1}^{n_i} C_{j-1}^{n_i} \lambda^{n_i-j+1} s^{j-1}\right)$ is a Hurwitz polynomial if $\hat{\hat{b}}_i/\hat{b}_i \geq 0.5+\varepsilon_i$, where ε_i is an arbitrarily positive number, $i=1,\dots,m$.

The existence of such a control law immediately follows the Assertion, and

$$u = H^{-1}(\hat{T}+T)x \tag{7}$$

III. STABILIZATION OF MULTI-INPUT CONTROLLABLE LINEAR SYSTEMS WITH PARAMETERS ASSUMING VALUES IN A COMPACT SUBSET OF PARAMETER SPACE

The results of II are now generalized to systems with parameters assuming values in a compact subset in the parameter space. The problem considered is more involved than that of previous section, since there are now infinite number of Hurwitz polynomials to be considered.

Assertion: For the multi-input controllable linear systems (1) of the form (2) with parameters assuming values in C_p, there always exists a control law which is linear state feedback and which can stabilize the systems having any parameter values in C_p if $b_{im} \geq \varepsilon > 0$, or $b_{iM} \leq \varepsilon < 0$, where ε is an arbitrarily small positive number.

Proof:

Let us consider a particular element $\Omega_1 \in C_p$, having the specific value for b_i

$$\Omega_1 = \left\{ a^{*k}_{ij} \in A^{k}_{ij} ,\ \begin{matrix} i,j=1,\cdots,m \\ k=1,\cdots,n_j \end{matrix} ,\quad b_i^* = \begin{matrix} b_{im} , & b_{im} > 0 \\ b_{iM} , & b_{iM} < 0 \end{matrix} ,\quad i=1,\cdots,m \right\}.$$

and any other element $\Omega \in C_p$.

For this pair of elements , we have

$$b_i / b_i^* \geq 1 > 0.5 + \varepsilon \quad , i=1,\cdots,m$$

Then from the results of II , we have shown the existence of a $\lambda^* > 0$ and of a control law $u = H^{-1}(\dot{T} + T)x$ with $\lambda \geq \lambda^*$, which stabilizes the systems having either of the parameter values. It is obvious that λ^* depends on the choice of $\Omega \in C_p$. Since the determination of $\lambda^*(\Omega)$ involves only continuous operation, the dependence of λ^* on $\Omega \in C_p$ is continuous. Since C_p is compact by assumption, the existence of a $\bar{\lambda}^*$ such that the control law $u = H^{-1}(\dot{T} + T)x$ with $\lambda \geq \bar{\lambda}^*$ which stabilizes the systems having any of the parameter values in C_p is guaranteed, where

$$\bar{\lambda}^* = \sup_{\Omega \in C_p} \left(\lambda^*(\Omega) \right)$$

Q.E.D.

REFERENCES

[1] W.M.Wonham,"On Pole Assignment in Multi-input Controllable Linear Systems",IEEE Trans. on Automatic Control, Vol.AC-12, December 1967.

[2] J.B.Pearson,and C.Y.Ding,"Compensator Design for Multivariable Linear Systems",IEEE Trans. on Automatic Control, Vol.AC-14,April 1969.

[3] H.H.Rosenbrock,"The Stability of Multivariable Systems",IEEE Trans. on Automatic Control, Vol.AC-17,Feburary 1972.

[4] D.E.Olson,and I.M.Horowitz,"Design of Dominate-type Control Systems with Large Parameter Variations",Int. J. of Control,Vol.12,1970.

[5] J.A.Rolnik,andI.M.Horowitz,"Feedback Control System Synthesis for Plants with Large Parameter Variations",IEEE Trans. on Automatic Control, Vol. AC-14,December 1969.

[6] D.G.Luenberger,"Canonical Forms for Linear Multivariable Systems,", IEEE Trans. on Automatic Control, Vol. AC-12,June, 1967.

[7] P.Brunovski," A Classification of Linear Controllable Systems", Kybernetika Cislo 3,Rocnik, June 1970.

[8] V.A.Chernyatin,"Stabilization of Linear Systems with Unknown Parameters ",Automation and Remote Control, Vol. 29, September 1968.

[9] C.Y.Ho,"Stabilization of Nonlinear Systems with Unknown Parameters by Linear Stste Feedback",Proc. of 15 th Midwest Symposium on Circuit Theory. May 1972.

DIGITAL COMPENSATOR SYNTHESIS

B. E. BONA
Autonetics Division
Rockwell International
3370 Miraloma Avenue
Anaheim, California 92803

ABSTRACT

Observers are used to synthesize a digital compensator for a single-input single-output sampled-data system. Closed-loop system poles are specified a priori and the digital compensator required to accomplish this pole setting is synthesized; the synthesis included systems with transport delay. The approach tends to bridge the gap between classical frequency domain, observer, and Kalman filter techniques.

1.0 INTRODUCTION

In designing control systems using modern techniques, the problem separates into two phases. In the first phase, one assumes that all state variables are accessible and closes the loop to satisfy some performance criterion. In the second phase, one proceeds to estimate or approximate those variables that are not directly accessible. The first phase is usually referred to as the control problem and the second phase is the estimation problem. Kalman filters [1] and Luenberger observers [2] are extensively used for the estimation problem.

Application [3, 4, 5] of these modern methods appears to have focused on time-varying, stochastic, multivariable, and other esoteric systems. Application to mundane constant, single-input single-output systems seems to be limited [6, 7, 8]. It is the intent of this paper to illustrate use of these modern methods to design a compensator for a sampled-date system. In the process, we tend to bridge the gap between classical root-locus, frequency, observer, and Kalman methods.

In the design approach taken, the plant poles are set first, assuming all state variables are accessible. The poles are set to achieve some specified bandwidth and/or time response. An observer is then designed to estimate the state variables. The poles of the oberver are placed at locations to achieve the desired closed-loop plant/observer response. By combining the control law and the observer, the compensator is synthesized to achieve the aforementioned pole setting.

This approach to compensator design is enlightening and leads to a flexibility [6, 8] in digital compensator design heretofore unexplored. Moreover, it lends insight into the more esoteric Kalman problems, where time simulation [9, 10] appears to be the only way to evaluate a closed-loop system. To help gain insight into closed-loop Kalman systems, the discussion is limited to unity observers [2]. The extension to minimum order observers is straightforward.

2.0 PLANT-OBSERVER MODEL

The problem is to synthesize a digital compensator for a single-input single-output system. The time invariant system (plant) is described by the state equations.

$$\dot{x} = Ax + Bu \tag{1}$$

$$y = m^T x \tag{2}$$

where x is an n-vector, u and y are scalars. Without further reference, we assume that the plant is sufficiently observable and controllable. Figure 1 graphically depicts (1) and (2) and further illustrates the nature of the problem. The output y is sampled at the constant sampling time $t = T_S$; the output of the digital compensator is the control variable u which passes through a zero-order hold.

The state equations for the sampled data system can be written as

$$x_{k+1} = \Phi x_k + \Delta u_k \tag{3}$$

$$y_k = m^T x_k \tag{4}$$

where

$$\Phi(T) = \Phi = \exp(AT_s) \tag{5}$$

$$\Delta(T) = \Delta = \int_0^{T_s} \Phi(t)Bdt \tag{6}$$

To keep the discussion simple, we assume the observer model and the plant are identical. Practically speaking, this means that the plant model described by (1) and (2) is the most significant and dominant part of perhaps a larger dimensional plant. The observer model is then based on the significant part of the plant. The rationale is that if the most significant part of the plant can be observed and controlled, then the overall plant will respond very closely to the modeled plant.

With the foregoing in mind, we consider the observer* [10] model

$$\hat{x}_k = \hat{x}'_k + b[y_k - \hat{y}'_k] \tag{7}$$

where

$$\begin{aligned} \hat{x}'_k &= \Phi \hat{x}_{k-1} + \Delta u_{k-1} \\ \hat{y}'_k &= m^T \hat{x}'_k \end{aligned} \tag{8}$$

In (7), y_k is the current sample of the output y, b is the observer gain vector, $\hat{x}_k$ is the estimate of x_k at time t_k and $\hat{x}'_k$ is the predicted or extrapolated estimate at time t_k. The observer equations can also be written in terms of the predicted or extrapolated estimate as follows. Using (7) and (8)

$$\begin{aligned} \hat{x}'_{k+1} &= \Phi \hat{x}_k + \Delta u_k \\ &= \Phi[(I-bm^T)\hat{x}'_k + by_k] + \Delta u_k \end{aligned} \tag{9}$$

Now consider the error

$$x_{k+1} - \hat{x}_{k+1} = \tilde{x}_{k+1} \tag{10}$$

Using (3) and (7) we arrive at the error equation

$$\tilde{x}_{k+1} = (I-bm^T)\Phi \tilde{x}_k \tag{11}$$

---

*For simplicity we consider the identity observer. The reader can easily extend the ideas to minimal order observers.

If b is chosen so that (11) is asymptotically stable, the observer estimates are unbiased. The estimates are used to control the system, i.e.,

$$u_k = -\alpha^T \hat{x}_k \tag{12}$$

The controlled plant (3) then takes the form

$$\begin{aligned} x_{k+1} &= \Phi x_k - \Delta\alpha^T\hat{x}_k \\ &= (\Phi - \Delta\alpha^T)\,x_k + \Delta\alpha^T\tilde{x}_k \end{aligned} \tag{13}$$

If (11) and (13) are combined we have the interesting result, obtained by Luenberger [2]

$$\begin{bmatrix} x_{k+1} \\ \hline \tilde{x}_{k+1} \end{bmatrix} = \left[\begin{array}{c|c} \Phi - \Delta\alpha^T & \Delta\alpha^T \\ \hline 0 & (I-bm^T)\Phi \end{array}\right] \begin{bmatrix} x_k \\ \hline \tilde{x}_k \end{bmatrix} \tag{14}$$

Equation (14) shows that the closed loop system (plant-observer) has a characteristic equation given by

$$F(z) = \left| zI - (\Phi - \Delta\alpha^T) \right| \; \left| zI - (I-bm^T)\Phi \right| \tag{15}$$

Inspection of (15) and (13) suggests that the plant poles are set with α, as though all state variables were directly available. Equation (15) also illustrates that the poles of the observer have been added to the closed loop system. The poles of the observer are set with the gain b.

3.0 COMPENSATOR

The compensator transfer function is easily derived from the equations in Section 2. Starting with (7), appropriately time shifted, leads to

$$\hat{x}_{k+1} = (I-bm^T)\,\hat{x}'_{k+1} + by_{k+1} \tag{16}$$

Now since

$$u_k = -\alpha^T \hat{x}_k$$

it follows from (8) that

$$\hat{x}'_{k+1} = (\Phi - \Delta\alpha^T)\,\hat{x}_k \tag{17}$$

so that

$$\hat{x}_{k+1} = (I-bm^T)(\Phi - \Delta\alpha^T)\,\hat{x}_k + by_k \tag{18}$$

Equation (18) is the closed loop observer; it describes the compensator dynamics. If we define

$$\Phi_c = (I-bm^T)(\Phi - \Delta\alpha^T) \tag{19}$$

as the compensator matrix, then on taking the z transform of (18) we have

$$\hat{x}(z) = (Iz - \Phi_c)^{-1}\, b\, z\, y(z)$$

so that the transfer function of the compensator is

$$\frac{u(z)}{y(z)} = \alpha^T (Iz - \Phi_c)^{-1}\, b\, z \tag{20}$$

Figure 2 illustrates the basic structure of the plant-observer pair. Note the explicit feedback of the control signal to the observer. This observer feedback is necessary in order that the plant-observer characteristic equation factor as in (15).

4.0 TRANSPORT DELAY

We now consider a compensator for systems with transport delay [8] caused by analog-to-digital conversion, data processing, and so on. The delay effectively causes the data y_k, y_{k+1}, ... to be delayed in time by some fraction, say, aT of T, where $0 \leq a \leq 1$. Figure 3 depicts the situation more vividly. The top scale indicates the time when the sampling takes place. The second scale indicates the transport delay. For example, it takes aT seconds after time t_k before y_k is converted and processed to arrive at a new control u_{k+a}. The last line of Figure 3 indicates the control times.

With reference to Figure 3, we can write (3), between control times, as

$$x_{k+1+a} = \Phi x_{k+a} + \Delta u_{k+a} \tag{21}$$

The control u_{k+a} is now given by

$$u_{k+a} = -\alpha^T \hat{x}'_{k+a} \tag{22}$$

where $\hat{x}'_{k+a}$ is the predicted or extrapolated value of x at time $t_k + aT$.

If (22) is substituted into (21), we arrive at

$$x_{k+1+a} = (\Phi - \Delta\alpha^T)\, x_{k+a} + \Delta\alpha^T\, [x_{k+a} - \hat{x}'_{k+a}]$$

$$= \Phi_p\, x_{k+a} + \Delta\alpha^T\, \tilde{x}_{k+a} \tag{23}$$

where

$$\tilde{x}_{k+a} = x_{k+a} - \hat{x}'_{k+a} \tag{24}$$

The predicted value of x at time $t_k + aT$ is

$$\hat{x}'_{k+a} = \Phi(aT)\, \hat{x}_k + \Delta(aT)\, u_{k-1+a} \tag{25}$$

where $\hat{x}_k$ is given by (7).

Figure 3 suggests that we can write

$$\hat{x}'_k = \Phi\,[(1-a)T]\, \hat{x}'_{k-1+a} + \Delta\,[(1-a)T]\, u_{k-1+a} \tag{26}$$

so that upon substituting (7) and (26) into (25) we get

$$\hat{x}'_{k+a} = \Phi(aT)\,(I-bm^T)\left\{\Phi\,[(1-a)T]\,\hat{x}'_{k-1+a} + \Delta\,[(1-a)T]\,u_{k-1+a}\right\} + \Delta(aT)\,u_{k-1+a} + \Phi(aT)\,by_k \tag{27}$$

This equation is a generalization of (9).

Now according to (22) the control at time (k - 1+a)T is

$$u_{k-1+a} = -\alpha^T\,\hat{x}'_{k-1+a} \tag{28}$$

so that upon substituting (28) into (27) we get

$$\begin{aligned}\hat{x}'_{k+a} = {} & \Phi(aT)\,(I-bm^T)\left[\Phi[(a-a)T] - \Delta[(1-a)T]\,\alpha^T\right]\hat{x}'_{k-1+a} \\ & - \Delta(aT)\,\alpha^T\,\hat{x}'_{k-1+a} + \Phi(aT)\,b\,y_k\end{aligned} \tag{29}$$

After some manipulation, the error equation can be written as

$$\begin{aligned}\tilde{x}_{k+a} &= \Phi(aT)\,(I-bm^T)\,\Phi[(1-a)T]\,\tilde{x}_{k-1+a} \\ &= \Phi_o\,(aT)\,\tilde{x}_{k-1+a}\end{aligned} \tag{30}$$

where

$$\Phi_o(a) = \Phi(aT)\,(I-bm^T)\,\Phi(T)\,\Phi(-aT)$$

If (23) is appropriately time shifted and combined with (30) we arrive at

$$\begin{bmatrix}\tilde{x}_{k+a} \\ \hline x_{k+a}\end{bmatrix} = \left[\begin{array}{c|c}\Phi_p & \Delta\alpha^T \\ \hline 0 & \Phi_o(a)\end{array}\right]\begin{bmatrix}x_{k-1+a} \\ \hline \tilde{x}_{k-1+a}\end{bmatrix} \tag{31}$$

From (31) it is clear that the characteristic equation of the closed loop system with delay is given by the product of determinants

$$F(z) = \left|\,zI - \Phi_o(a)\,\right|\,\left|\,zI - \Phi_p\,\right| \tag{32}$$

Moreover, since $\Phi_o(a)$ is similar to $\Phi_o(o)$, it follows that the characteristic equation (32) is identical to (15), for no transport delay. Thus, transport delay can be accounted for exactly by using the compensator defined by (29) and (22).

If we write (29) as

$$\hat{x}'_{k+a} = \Phi_c(a)\,\hat{x}'_{k-1+a} + b\,(a)\,y_k \tag{33}$$

where

$$\Phi_c(a) \triangleq L(a) \quad \Phi_p + [L(a) - I] \, \Delta(aT) \, \alpha^T \tag{34}$$

$$\begin{aligned} L(a) &\triangleq \Phi(aT) \, (I - bm^T) \, \Phi(-aT) \\ b(a) &= \Phi(aT) \, b \end{aligned} \tag{35}$$

then the transfer function of the compensator can be written as

$$\frac{u(z,a)}{y(z)} = \alpha^T [zI - \Phi_c(a)]^{-1} \, z \, b(a) \qquad 0 \le a < 1 \tag{36}$$

which is a generalization of (20).

5.0 DESIGN PROCEDURE

The design procedure can be summarized in the following steps.

(1) Specify the closed-loop plant poles (13) and determine the vector α .

(2) Specify the observer poles (11) and determine the vector b.
The closed-loop system will now have the characteristic equation given by (15).

(3) Using the vectors α and b obtained in steps (1) and (2) determine the compensator (20) for no transport delay or (36) for transport delay.

To illustrate the design procedure consider the system shown in Figure 4. Neglecting the control disturbance at the input and the noise disturbance on the output, the system equations can be written as

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} u_k \tag{37}$$

$$y_k = [1 \quad 0] \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}$$

To stabilize the plant let

$$u_k = - \alpha^T \hat{x}_k$$

where

$$\alpha^T = \begin{bmatrix} \frac{\alpha_1}{T^2} & \frac{\alpha_2}{T} \end{bmatrix}$$

Then the characteristic equation (see 15) of the controlled plant is

$$z^2 - \left(2 - \alpha_1 - \frac{\alpha_2}{2}\right) z + \left(1 - \alpha_2 + \frac{\alpha_1}{2}\right)$$

so that α_1 and α_2 can be chosen to set the plant poles at desired locations. The observer transition matrix is

$$(I-bm^T)\,\Phi = \begin{bmatrix} 1-b_1 & (1-b_1)T \\ -b_2 & 1-b_2T \end{bmatrix} \tag{38}$$

and its characteristic equation is

$$z^2 - (2-b_1-b_2T)\,z + (1-b_1)$$

For simplicity, let us choose the fastest time observer. This is done by setting the observer eigenvalues at the origin. Thus $b_1 = 1$ and $b_2 = \frac{1}{T}$. The observer without feedback is

$$\hat{x}_{k+1} = \begin{bmatrix} 0 & 0 \\ -\frac{1}{T} & 0 \end{bmatrix} \hat{x}_k + \begin{bmatrix} 1 \\ \frac{1}{T} \end{bmatrix} y_k \tag{39}$$

On taking the z transform of (39), we have

$$\hat{x}_1\,(z) = y\,(z) \tag{40}$$

$$\hat{x}_2\,(z) = \frac{z-1}{T\,z}\,y\,(z) \tag{41}$$

Equations (40) and (41) are the traditional transfer functions used to determine position and rate, respectively. Conventional procedure is to feed the appropriate linear combination of these variables back to stabilize the system. Root locus or other means is then used to ascertain the location of the closed-loop poles. The observer approach, on the other hand, leads to known closed-loop pole locations, providing the feedback loop to the observer is closed, as indicated in Figure 2 and (17). Thus, if feedback to the observer is provided (39) is modified according to (18) so that

$$\hat{x}_{k+1} = \begin{bmatrix} 0 & 0 \\ -\frac{1}{T}\left[1+\frac{\alpha_1}{2}\right] & -\frac{\alpha_2}{2} \end{bmatrix} \tilde{x}_k + \begin{bmatrix} 1 \\ \frac{1}{T} \end{bmatrix} u_k \tag{42}$$

The transfer functions are

$$\hat{x}_1\,(z) = y\,(z)$$

$$\hat{x}_2\,(z) = \frac{z-(1+\alpha 1/2)}{T\,(z+\alpha 2/2)}\;y\,(z) \tag{43}$$

Comparison of (41) and (43) shows how the compensator poles and zeros are moved when the loop is closed. If the compensator defined by (43) is used, the closed-loop poles will be determined by the poles of the plant and the observer in accordance with (15). If the compensator defined by (41) is used, the locations of closed loop are undetermined.

It should be clear from the above example that the approach afforded by the plant-observer pair requires that the designer specify closed-loop pole locations. This information is generally known either in frequency-domain specifications and/or in time-response specifications. Consider, for example, the two-integrator system. This system represents a model of a rotating inertia where x_1 is angle and x_2 is rate. The control signal u is used to drive a torquer motor to stabilize the loop. To prevent exciting mechanical resonance, it is desired to keep the loop response below some specified frequency. In addition to resonance, high frequency electronic noise is also a consideration. Thus, to suppress the high-frequency noise, it is desired to have low loop gain at some specified frequency.

Using the observer design approach, we first set the closed-loop plant poles (e.g., poles of $\Phi - \Delta \alpha^T$) to achieve the low frequency response. Figure 5 depicts the desired pole locations in the s-plane. The next step is to set the high frequency poles; this is done by setting the poles of the observer (i.e., poles of $(I - bm^T)\Phi$) as indicated in Figure 5. These pole settings determine both gain vectors α and b, so that the compensator is specified by (20). Compensator poles and zeros are indicated in Figure 5 to complete the picture for determining frequency and time response. On Figure 5, we have sketched in some trajectories from open-loop poles to closed-loop poles in root locus tradition. Intuitively, we show the loci of the closed-loop plant poles starting on the open-loop plant poles and ending at their closed-loop locations. The remaining poles start on the compensator poles and end on the observer poles (i.e., on poles of $(I - bm^T)\Phi$). The above interpretation of plant-observer-compensator poles allows the designer to relate frequency domain specifications time response.

If system (37) has transport delay, we follow the preceding steps only now (36) is used to determine the compensator. To be specific let $\alpha_1 = 3/4$ and $\alpha_2 = 5/8$, then the plant poles are at $1/2 \pm j\sqrt{2}/2$. The closed-loop system would have the characteristic equation

$$(z^2 - z + 3/4)\, z^2$$

To determine the compensator, we need the following matrices

$$\Phi[(1-a)T] = \begin{bmatrix} 1 & (1-a)T \\ 0 & 1 \end{bmatrix}$$

$$\Delta[(1-a)T] = \begin{bmatrix} (1-a)^2 \dfrac{T^2}{2} \\ (1-a)\, T \end{bmatrix}$$

and similar matrices for $\Phi(aT)$ and $\Delta(aT)$. Equation (36) is now used to determine the required compensator. Table 1 gives the pole locations for typical compensator designs when a = 0, a = 1/2 and a = 1.

Table 1. Poles of Compensator (Two-Integrator Plant)

Time delay = a	Poles
0	0, -.3
1/2	-.5, -.26
1	.5 ± .65

To continue the example, suppose there is a requirement that the system follow a step input disturbance without error. In terms of classical control systems, the requirement would call for a type I servo, or for one with infinite d-c stiffness. In terms of modern estimation theory, the requirement would call for a Kalman filter that would estimate random constants. To design the required compensator, we include the step input disturbance in an augmented system model. For the example at hand, the augmented system model becomes

$$\begin{bmatrix} x_1\ (k+1) \\ x_2\ (k+1) \\ x_3\ (k+1) \end{bmatrix} = \begin{bmatrix} 1 & T & \frac{T^2}{2} \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1\ (k) \\ x_2\ (k) \\ x_3\ (k) \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \\ 0 \end{bmatrix} u_k \tag{44}$$

$$y_k = (1 \quad 0 \quad 0)\ x_k$$

Note that (44) is observable, but uncontrollable. Thus we can design an observer to estimate state x_3, the step disturbance, but we cannot control state x_3. We can, however, apply a control signal to nullify its effect. To see this let

$$\alpha^T = \begin{bmatrix} \frac{\alpha_1}{T^2} & \frac{\alpha_2}{T} & \alpha_3 \end{bmatrix} \tag{45}$$

then

$$\Phi_p = (\Phi - \Delta\alpha^T) = \begin{bmatrix} 1-\frac{\alpha_1}{2} & T(1-\frac{\alpha_2}{2}) & \frac{T^2}{2}(1-\alpha_3) \\ -\frac{\alpha_1}{T} & 1-\alpha_2 & T\ (1-\alpha_3) \\ 0 & 0 & 1 \end{bmatrix} \tag{46}$$

with characteristic equation

$$(z-1) \left\{ z^2 - (2-\alpha_1 - \frac{\alpha_2}{2})\ z + 1 - \alpha_2 + \frac{\alpha_1}{2} \right\} \tag{47}$$

Equation (47) shows that the closed-loop poles are independent of α_3, but for infinite stiffness it is necessary that the compensator have a pole at $z = +1$. It is easy to verify that if $\alpha_3 = 1$, the compensator matrix

$$\Phi_c = (I - bm^T)\ \Phi_p$$

will have the required pole. Thus for infinite stiffness, it is necessary to feedback $\hat{x}_3$ with unity gain.

To complete the example, consider the noise disturbance on the output. To be explicit, suppose the disturbance is exponentially correlated noise with correlation frequency β. Then we can write

$$\dot{\eta} = -\beta\eta + \omega$$

where ω is white noise. If we were designing a Kalman filter, it would be necessary to assume an explicit value for the variance of ω. In designing the observer, we do not require an explicit variance, but in setting the observer poles, we tacitly make some assumption regarding its value. Clearly, we would not want the fastest observer (see (39)) if the noise were large.

In designing the compensator for noise on the output, we again form an augmented system. If we neglect the disturbance on the input, the augmented plant model is

$$\begin{bmatrix} x_1\,(k+1) \\ x_2\,(k+1) \\ \eta\,(k+1) \end{bmatrix} = \begin{bmatrix} 1 & T & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-\beta T} \end{bmatrix} \begin{bmatrix} x_1\,(k) \\ x_2\,(k) \\ \eta(k) \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \\ 0 \end{bmatrix} u_k \tag{49}$$

$$y_k = (1 \quad 0 \quad 1)\; x_k$$

Note that (49) is observable, but uncontrollable. Thus we can estimate the noise variable, but cannot explicitly control it. To see what is accomplished by including it in the model, suppose we compute the zeros of the compensator. By repeated use of the matrix identity (x and y vectors)

$$\left[I + (SI-A)^{-1}\, x\, y^T \right]^{-1} = I - \frac{(SI-A)^{-1}\, xy^T}{1 + y^T (SI-A)^{-1} x} \tag{50}$$

we can rewrite (20) as

$$\frac{u(z)}{z\, y(z)} = \alpha^T \frac{\text{adj}\,(zI-\Phi)b}{|z-\Phi_c|} \tag{51}$$

where Φ is the system matrix in (49) and Φ_c is the compensator matrix defined in (20). For noise on the input, the matrix Φ can always be partitioned as

$$\Phi = \left[\begin{array}{c|c} \Phi_{11} & 0 \\ \hline 0 & \Phi_{22} \end{array}\right] \tag{52}$$

where Φ_{22} is the transition matrix of the noise process. In (49), Φ_{22} is the scalar $e^{-\beta T}$.

Now since we cannot control the noise, the feedback vector α can also be partitioned as $[\underline{\alpha}_1^T \mid 0]$. For the example at hand, we feedback $\hat{x}_1$ and $\hat{x}_2$, so we let

$$\alpha^T = [\underline{\alpha}_1^T \mid 0] \tag{53}$$

Now using (51) we can establish the result

$$\alpha^T \text{adj}\,(zI-\Phi)\, b = \underline{\alpha}_1^T \text{adj}\,(zI-\Phi_{11})\, \underline{b}_1\; |z-\Phi_{22}|$$

where $\underline{b}_1$ is a compatible partition. This equation shows that the compensator will have zeros given by the determinant $|z-\Phi_{22}|$, which, in turn, is the characteristic equation of the noise process (i.e., the poles of the noise become compensator zeros). For above example, we have

$$\alpha^T \text{adj}\,(zI-\Phi)b = \left\{ \alpha^T b\,(z-1) + \frac{b_2\,\alpha_1}{T} \right\} (z-e^{-\beta T})$$

If the noise is described by a slightly damped sinusoid, one could easily design a notch filter into the compensator.

The foregoing discussion applies equally well to Kalman filters [3]. If colored observation noise is included in the design, it would appear that zeros are deposited in the appropriate locations. Pole placement for the Kalman filter or for the observer is determined by the variance of the noise. For the Kalman filter one determines the required variances a priori and the poles (e.g., steady state poles) follow. For the observer, one sets the poles a priori, using his tacit knowledge of how noisy things are.

CONCLUSIONS

Modern control system methods provide a systematic approach for synthesizing feedback compensators. The observer design approach allows one to set closed-loop poles a priori and synthesize the required compensator. The design approach allows the classical designer to relate, and effectively marry, the classical design methods (and specifications) with the more modern methods. The observer approach is much more flexible than classical methods.

REFERENCES

(1) R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," Trans., ASME, J. Basic Eng., Ser. D, pp. 95-108, March 1961.

(2) D. G. Luenberger, "An introduction to observers," IEEE Trans. Automat. Contr., Vol AC-16, pp. 596-602, December 1971.

(3) B. E. Bona, "Application of observers and optimum filters to inertial systems," IFAC Sump. Multivariable Control Systems, Duesseldorf, Germany, 1968.

(4) L. Novak, "The design of an optimal observer for linear discrete-time dynamical systems," 4th Asilomar Conf. Circuits and Systems, 1970.

(5) Y. O. Yuksel and J. J. Bongiorno, "Observers for linear multivariable systems with applications," IEEE Trans. Automat. Contr., Vol AC-16 December 1971.

(6) B. E. Bona, "An adaptive digital compensator for saturating control systems," Sixth Asilomar Conference on Circuits and Systems, Pacific Grove, CA, November 1972.

(7) B. E. Bona, "Digital filter design using observers," Fifth Asilomar Conference on Circuits and Systems, Pacific Grove, CA, November 1971.

(8) B. E. Bona, "Digital compensators with transport delay," 1972 IEEE Conference on Decision and Control, New Orleans, La, December 1972.

(9) J. S. Mendel, "On the need for and use of a measure of state estimation errors in the design of quadratic-optimal control gains," IEEE Trans. on Automatic Control, Vol AC-16, No. 5, 1971.

(10) B. E. Bona, "Linear regulators and observers," IEEE Trans. on Automatic Control, Vol AC-18, No. 1, 1973.

(11) B. E. Bona, "Designing observers for time-varying state systems, Fourth Asilomar Conference on Circuits and Systems, Pacific Grove, CA, 1970.

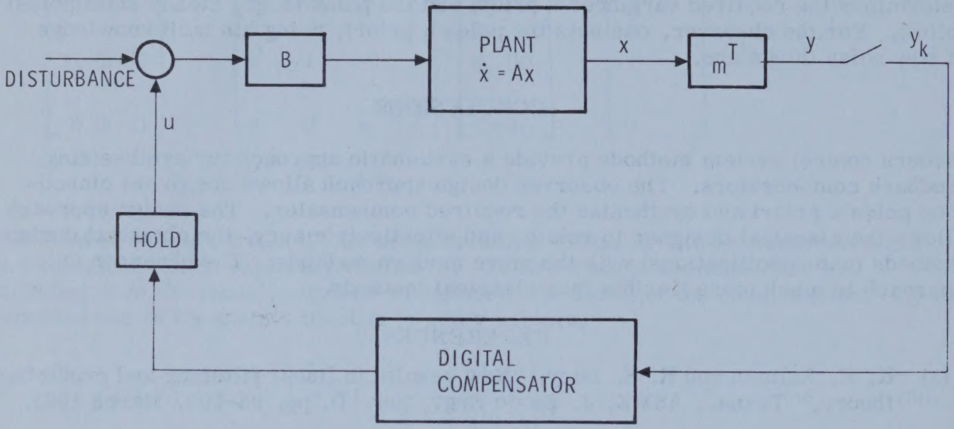

DISTURBANCE
B
PLANT
ẋ = Ax
x
m^T
y
y_k
u
HOLD
DIGITAL COMPENSATOR

Figure 1. Digital Compensation of Continuous Plant

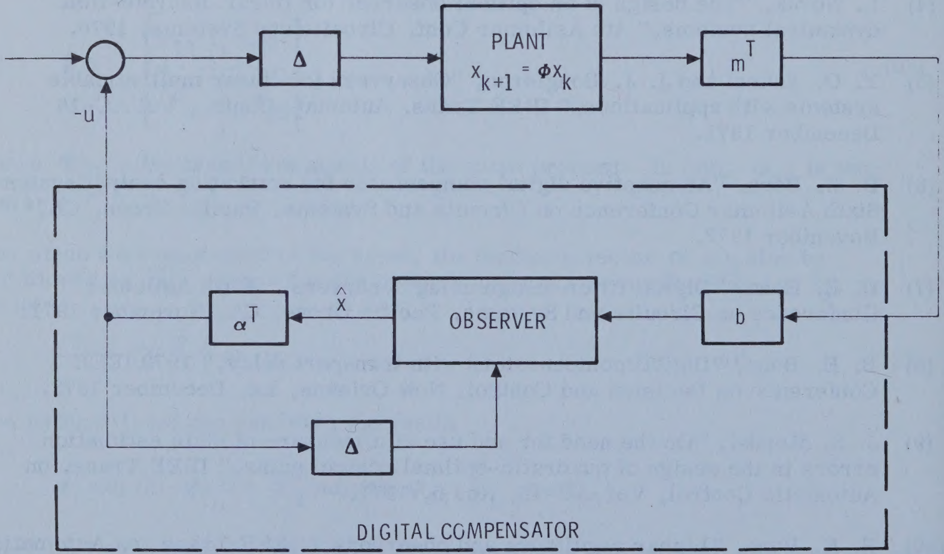

Δ
PLANT
x_{k+1} = Φx_k
m^T
-u
α^T
x̂
OBSERVER
b
Δ
DIGITAL COMPENSATOR

Figure 2. Structure of Plant-Observer Controlled System

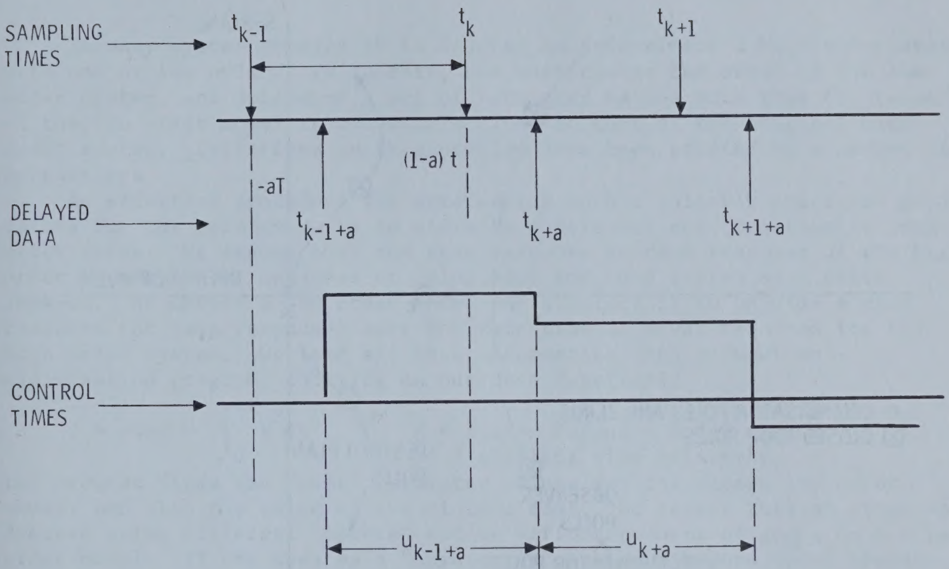

SAMPLING TIMES
t_{k-1}
t_k
t_{k+1}
$(1-a)\,t$
$-aT$
DELAYED DATA
t_{k-1+a}
t_{k+a}
t_{k+1+a}
CONTROL TIMES
u_{k-1+a}
u_{k+a}

Figure 3. Sampling and Control Reference Time

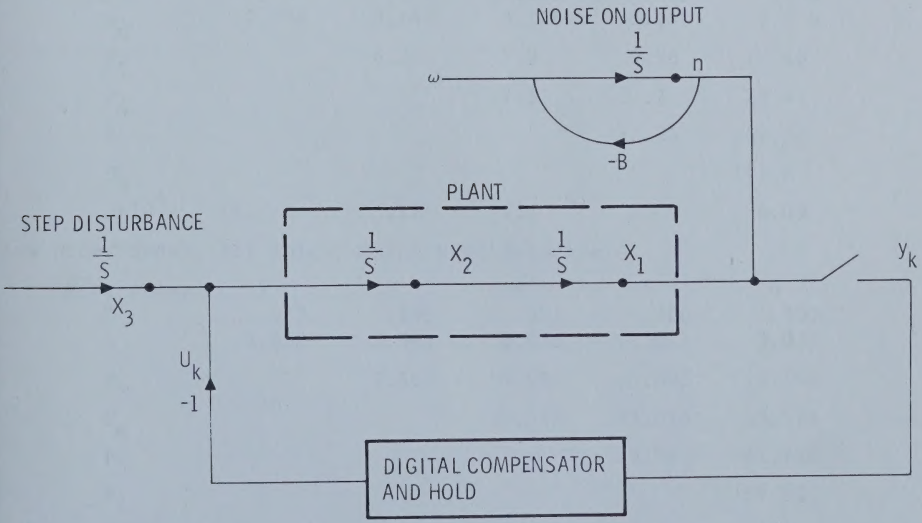

NOISE ON OUTPUT
$\frac{1}{S}$
ω
n
-B
PLANT
STEP DISTURBANCE
$\frac{1}{S}$
X_3
$\frac{1}{S}$
X_2
$\frac{1}{S}$
X_1
y_k
U_k
-1
DIGITAL COMPENSATOR AND HOLD

Figure 4. System for Example Design

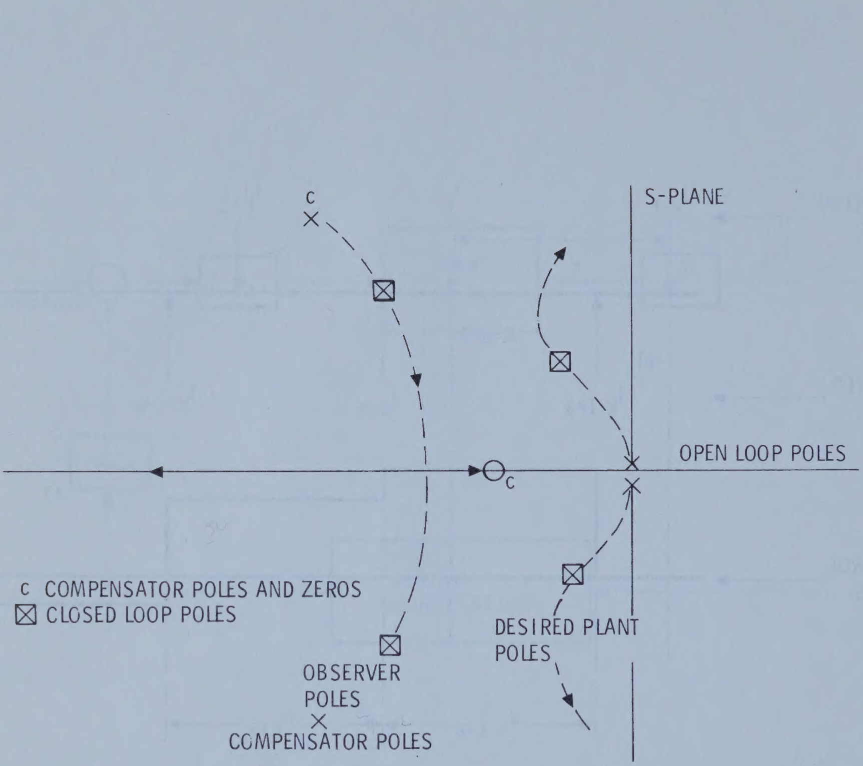

c
S-PLANE
OPEN LOOP POLES
c
c COMPENSATOR POLES AND ZEROS
☒ CLOSED LOOP POLES
DESIRED PLANT POLES
OBSERVER POLES
COMPENSATOR POLES

Figure 5. Pole-Zero Plot of Plant-Observer-Compensator

LOW ORDER MODELS FOR DYNAMIC SYSTEMS

J. CANTALAPEIDRA, Spanish Navy
N. V. LAM, South Vietnamese Navy
G. J. THALER, Naval Postgraduate School, Monterey, California

In many system studies it is desired to approximate a high order system with one of low order. To do this, one must choose the order of the low order system, and determine a set of parameter values such that the response of the low order model is acceptably close to that of the original high order system. Variations on this problem have been studied by a number of researchers

An effective procedure for determining both a suitable order and good values for the parameters is to minimize a selected cost function in parameter space. We assume that the step response or ramp response of the high order system can be measured or calculated and then stored as a table look-up. We choose a low order model and simulate it to provide a step response (or ramp response) over the same time interval required for the high order system. We feed all this information into a function minimization program, choosing as our cost functions:

$$J = \text{cost} = \int_0^{tf} E^2 dt \qquad E = \text{system output - model output}, \quad tf = \text{settling time of system.}$$

the program finds the "best" parameter values for the chosen low order model, and also the value of the minimum cost. We repeat this as often as desired using different "orders" and/or various numbers of zeros in our low order model. If one uses as a "high order" system a second order system with time delay, typical results are:

Low Order Model (All Poles) Using Step Response:

# of Poles	2	3	4	5	6
ζ	.35	.288	.297	.303	.309
ω_n	2.734	3.145	3.148	3.148	3.138
P_3		6.24	7.8	9.35	13.46
P_4			37.1	29.7	25.4
P_5				87.24	41.57
P_6					111.9
$J*10^4$	131	5.17	4.7	5.09	6.09

Low Order Model (All Poles) Using Ramp Response:

# of Poles	2	3	4	5	6
ζ	.432	.299	.305	.306	.303
ω_n	2.658	3.009	3.036	3.041	3.055
P_3		7.466	14.924	15.095	15.194
P_4			15.513	25.014	35.575
P_5				39.903	41.728
P_6					59.951
$J*10^4$	32	5.9	4.9	4.7	4.7

CONCLUSIONS: Second order models are not very good. A "best" order (and a "best" number of zeros) can be found.

AN EXPERIMENTAL STUDY OF A HYBRID ADAPTIVE CONTROL SYSTEM

E. F. LIZEWSKI AND R. V. MONOPOLI
Department of Electrical and Computer Engineering
University of Massachusetts
Amherst, Massachusetts

ABSTRACT

A Liapunov type model reference adaptive control system with five adjustable gains is implemented using a PDP-11 digital computer and an EAI 380 analog computer. The plant controlled is a laboratory type d-c servo system. It is made to follow closely a second order linear model. The experimental results demonstrate the feasibility of implementing this rather complex design using only a mini-computer and a reasonable number of operational amplifiers. Also, it points out that satisfactory performance can be achieved even when certain assumptions necessary for the theory are not satisfied.

INTRODUCTION

Many theoretical and computer simulation studies have been made on the synthesis of model reference adaptive control systems based on the application of Liapunov's second method [1]-[3]. The basic Liapunov design [4,5,8] assumes: the availability of, except for a restricted class of plants [6], all plant states or their generation by an observer or differentiators; the availability of an unbounded control input; a noiseless, disturbance free environment; constant or slowly varying parameters; and an input of sufficient frequency content to excite all plant states [7]. The accessibility of internal plant parameters may also be required [8].

Due at least partially to the above requirements, little has been done with the application of these design techniques to physical systems. Those which have been implemented are usually restricted to first order plants where an analog controller adapts for one or two unknown parameters [9,10].

This paper reports on the hybrid (analog-digital) implementation of a model reference adaptive control system which uses a modified Liapunov technique [11] for the adaptive controller design. The problem considered here is the feasibility of implementing this type design when many of the requirements imposed by the theory cannot be met in practice. The design procedure assumes the availability of a plant output derivative, an unbounded control input, constant parameters, and a noiseless, disturbance free environment. In this work a plant with nonlinear characteristics is forced to track a linear second order underdamped model in a region where the nonlinearities are pronounced. Adaptation occurs using an approximate plant output derivative and a saturating control input while operating in a noisy environment where the system is subject to parameter variations and disturbance inputs.

The general system layout is shown in Fig. 1. The objective is the velocity control, using plant input modification, of a small armature connected d-c motor. The model reference and approximate derivative circuits are implemented on an analog computer. A/D and D/A converters are used to close the adaptive loop through a mini-computer (PDP-11) used to implement the adaptive gain algorithms and the required signal multiplications.

THEORY AS APPLIED TO THE GENERAL, IDEAL SYSTEM

The plant is described by the differential equation

$$\ddot{\omega}_p + a_{ip}\dot{\omega} + a_{2p}\omega_p + d_1 f_1(\omega_p) + d_2 f_2(\omega_p) = K_p v + K_u u \tag{1}$$

where ω_p is the angular velocity of the motor shaft, v and u are the reference and control input, the forms of the coulomb friction and stiction friction $f_1(\omega_p)$ and $f_2(\omega_p)$ are assumed known, and coefficients a_{1p}, a_{2p}, d_1, d_2, K, and K_2 are assumed constant and unknown. The control input is taken to be $u = \underline{k}^T\underline{z}$ where $\underline{k}$ is the vector of adaptive gains, $\underline{k}^T = [k_1(t), k_2(t)],\ldots k_5(t)]$, and $\underline{z}^T = [v, \dot{\omega}_p, \omega_p, f_1(\omega_p), f_2(\omega_p)]$. The model reference is described by

$$\ddot{\omega}_d + a_{1d}\dot{\omega}_d + a_{2d}\omega_d = K_d v \tag{2}$$

where ω_d, the model output, is the desired angular velocity of the plant, and the model is stable ($a_{1d} > o$, $a_{2d} > o$). Subtracting (1) from (2) and using the expression for u gives the error equation

$$\ddot{e} + a_{1d}\dot{e} + a_{2d}e = \underline{z}^T\underline{\Delta} \tag{3}$$

where $e = \omega_d - \omega_p$, $\underline{\Delta}^T = [\Delta_1, \ldots \Delta_5] = [(k_d - K_p - K_u k_1(t)), (a_{1p} - a_{1d} - K_u k_2(t)), (a_{2p} - a_{2d} - K_u k_3(t)), (d_1 - K_u k_4(t)), (d_2 - K_u k_5(t))]$. The Liapunov function used is

$$V = (1/2)(\underline{e}^T P\underline{e} + (\underline{\Delta} + \beta\ \underline{z}\ \underline{p}^T\underline{e})^T\ \lambda(\underline{\Delta} + \beta\ \underline{z}\ \underline{p}^T\underline{e}) \tag{4}$$

where $\underline{e}^T = [\dot{e}, e]$, λ and β are 5 x 5 diagonal matrices with elements $\lambda_i > o$, $\beta_i > o$, P is a 2 x 2 symmetric matrix, and $\underline{p}^T = [p_{11}, p_{22}]$ where p_{11}, p_{22} are the diagonal elements of P. With the adaptive gain algorithm taken to be $\dot{\underline{k}} = (1/K_2)(\lambda^{-1}\underline{z}\ \underline{p}^T\underline{e} + \beta(d/dt)(\underline{z}\ \underline{p}^T\underline{e}))$, the time derivative of V becomes $\dot{V} = -e^T Q\underline{e} - \underline{p}^T\underline{e}\ \underline{e}^T\underline{p}\ \underline{z}^T\beta\ \underline{z}$, where $Q = A^T P + PA$, $(A = [\underline{a}_1, \underline{a}_2]$, $\underline{a}_1^T = [o, -a_{2d}]$, $\underline{a}_2^T = [1, -a_{1d}])$, is chosen to be positive definite yielding a positive definite P. $\dot{V}$ is negative definite in $\underline{e}$, hence $\lim_{t\to\infty} \underline{e} = \underline{o}$.

The basic Liapunov design [8] is obtained with $\beta = [o]$. The modified design ($\beta \neq [o]$) results in direct feedback of the error states to the adaptive gains and adds an additional negative definite term to $\dot{V}$ (the second term). It was shown in [11] by simulation that this modified design results in improved convergence and damping of the error response. Experimental confirmation of these results is demonstrated herein.

THEORY AS APPLIED TO THE ACTUAL SYSTEM

The plant, shown in Figure 2, includes noise (and/or disturbance) input n and a low pass filter to eliminate some of the tachometer brush noise before taking the derivative of signal $\hat{\omega}_p$. The nonlinear frictional torque characteristics, modeled as in [12], are $f_1(\omega_p) = \mathrm{sgn}(\omega_p)$ and $f_2(\omega_p) = \mathrm{sgn}(\omega_p)/(\gamma(\omega_p) + \varepsilon)$ where γ and ε are positive constants. The low pass filter is now considered part of the plant and $\hat{\omega}_p$ rather than ω_p is the "plant output". The system error equation written on this basis (i.e. taking $e = \omega_d - \hat{\omega}_p$) and including n is

$$\ddot{e} + a_{1d}\dot{e} + a_{2d}e = \underline{\hat{z}}^T\underline{\hat{\Delta}} + \underline{\ell}^T\underline{w} \tag{5}$$

where $\underline{\hat{z}}^T = [v, \dot{\hat{\omega}}_p, \hat{\omega}_p, f_1(\hat{\omega}_p), f_2(\hat{\omega}_p)]$, $\underline{\hat{\Delta}}$ has the elements of $\underline{\Delta}$ except for d_1 and d_2 which now are nonlinear functions of $\hat{\omega}_p$, $\dot{\hat{\omega}}_p$, and n. $\underline{\ell}$ is a 7 x 1 vector of gains, and $\underline{w}$ is a 7 x 1 vector with elements which are complicated functions of n, $\dot{n}$, $\ddot{n}$, $\hat{\omega}_p$, $\dot{\hat{\omega}}_p$, $\ddot{\hat{\omega}}_p$, $\dddot{\hat{\omega}}_p$, $\partial f_1(\omega_p)/\partial\omega_p$, and $\partial f_2(\omega_p)/\partial\omega_p$. These elements are not available for control. Only the elements of $\underline{\hat{z}}$ can be used to generate the control input $\hat{u} = \underline{\hat{k}}^T\underline{\hat{z}}$ where now the adjustable gains are $\dot{\underline{\hat{k}}} = (1/K_2)(\lambda^{-1}\underline{z}\,\underline{p}^T\underline{e} + \beta\,(d/dt)(\underline{\hat{z}}\,\underline{p}^T\underline{e}))$. With this algorithm, the derivative of the Liapunov function evaluated for (5) is

$\dot{V} = -\underline{e}^T Q\underline{e} - \underline{p}^T\underline{e}\,\underline{e}^T\underline{p}\,\underline{\hat{z}}\,\beta\,\hat{z} + \underline{p}^T\underline{e}\,\underline{\ell}^T\underline{w} + \dot{\underline{\hat{\Delta}}}_1^T\lambda(\underline{\hat{\Delta}} + \beta\,\underline{\hat{z}}\,\underline{p}^T\underline{e})$, where $\dot{\underline{\hat{\Delta}}}_1$ contains time derivatives of terms in $\underline{\hat{\Delta}}$ other than the time derivatives of the adaptive gains. The last two terms in $\dot{V}$, present because of the non-ideal conditions (i.e. noise and the low pass filter) are indefinite. Therefore it is no longer possible to guarantee that $\underline{e} \to \underline{o}$. However, use of the modified design ($\beta \neq [o]$) is shown experimentally to be of considerable help in overriding these indefinite terms by contributing the negative definite term involving β to $\dot{V}$.

SYSTEM IMPLEMENTATION

Fig. 3 shows the overall equipment layout. The model reference circuit and plant output derivative circuit are patched on the analog computer. The derivative circuit has the transfer function 100s/(s + 100). Over the significant frequencies encountered it is capable, depending on the noise level, of generating the plant acceleration with approximately 95% to 75% accuracy. The model-plant error is taken as the difference between the filtered tach voltage and the model output voltage. It was found necessary to use a filter time constant of 0.5 sec. (Fig. 4 shows the noisy nature of the derivative even for this large value of τ). Therefore, there is a significant lag between the actual plant output and that of the model. (An approach to overcoming this lag problem would be the use of an equivalent filter on the model output).

A mini-computer is used to implement the adaptive laws and generate the control input. Over a one-millisecond period all plant and model variable input channels are sampled and the portion of the control input associated with one of the adaptive gains is calculated; i.e., say $u_{k_1} = k_1 v$. This portion of the control input is then updated. A zero input channel is also sampled every millisecond to set the zero voltage bias level. The five element control input is therefore completely updated every five milliseconds. A trapezoidal integration over this five millisecond time base is used to calculate the integral term portions of the adaptive gains. With a plant time constant of approximately 0.3 sec., errors due to the trapezoidal approximation and discretization are safely ignored. The control input magnitude is bounded by the D/A converter to ± 5 volts. Since 5 volts saturates the servo power amplifier, gain K_u is used to attenuate the control input.

Many scaling operations, absorbed by the arbitrary gains λ_i and β_1 on the adaptive laws, are made necessary by the computer word size and integer math hardware used (floating point software is much too slow for real time use here). This residue modulo 8 integer package results in truncation error; i.e., $1/5 = 0$ with a remainder of 1 while $7/5 = 1$ with a remainder of 2. (The effects of this truncation error on performance were not investigated in detail; however, it appears to be one of the principle sources of the steady state error level observed.)

PERFORMANCE

A low damping ratio for the model was intentionally chosen, since this represents, in a sense, a "worst case". The model reference damping ratio and natural frequency were set at 0.25 and 5.8 r/s. When tracking a square wave reference input this low damping ratio caused appreciable saturation of the control input in the D/A converter. When the gain K_u was increased in an attempt to avoid saturation in the D/A, performance deteriorated due to a high noise level and/or saturation of the servo power amplifier.

Each adaptive gain is comprised of an integral term and a proportional term; i.e., $k_i = (1/\lambda_i)\int(*)dt + \beta_i(*)$. The notation $\forall I + \forall P$ means that all five adaptive gains are used to generate the control input and each gain is comprised of both the integral term and the proportional term. $Ik_ik_j + Pk_1k_m$ means the control input is generated using the integral terms of k_i and k_j and the proportional terms of k_1 and k_m. IEM denotes the integral of the error magnitude over the first 30 sec. of adaptation.

Figs. 5 through 9 correspond to a 0.3 Hz, ± 1 volt square wave reference input. Fig. 5 shows the plant and model responses to this reference input with no adaptive control input applied. The extremely poor plant response is due to the relatively large effect of various system nonlinearities when operating with the fairly low level reference input. Some of these nonlinearities could not be included in the controller design. Adaptation is begun with non-zero initial conditions. Fig. 6 shows the response using $\forall I + \forall P$. The arbitrary gains $1/\lambda_i$ and β_i are moderate. When these gains are large, stability and sensitivity to noise become critical, but convergence time can be brought below 1 sec. Fig. 7 is the control input, u, corresponding to Fig. 6. The shaded region represents the high frequency contribution of the proportional terms. An error of 2 r/s coinciding with the saturation of u occurs during overshoot.

Fig. 8 is the response using ∇I; i.e., $[\beta] = [0]$. Here convergence is much slower and an error of 7 r/s occurs during overshoot. Fig. 9 is the response using ∇P $[\lambda]^{-1} = [0]$; it is seen that the proportional terms alone are sufficient to hold performance within reasonable limits. In all cases there is considerable saturation of u. With the square wave input, where there is considerable acceleration through zero velocity, the gains used to compensate for the nonlinear frictional characteristics, k_4 and k_5, don't make a significant contribution to performance, as illustrated in Table 1. Also, the integral term of the gain associated with the noisy plant derivative, Ik_2, does not contribute substantially.

Fig. 10 is the response to a 0.3 HZ, ± 1 volt sine wave reference input using ∇I. Here IEM = 54 r. Compared to the square wave, these results are considerably better. This is due to less saturation of u and a less noisy derivative because of less tachometer brush bounce. In this case Ik_2 contributes appreciably to performance.

To test performance where the friction has its greatest effect, a "one sided" 0.3 Hz, 0.2 to 1.2 volt square wave and sine wave reference inputs were used. The performance was qualitatively similar to that found above except different adaptive gains are of importance here where the acceleration through zero velocity is negligible. Table 2 gives the IEM's for the square wave case. The results indicate the gain associated with the coulomb friction, k_4, is of primary importance in compensating for the nonlinear frictional characteristics. Again, Ik_2 does not contribute when tracking the square wave.

It was found that the integral terms diverged and approached the computer word size, causing the system to go unstable. This occurred in 20. min. with the sine wave and in 110. min. with the square wave. The integrals did not diverge, but oscillated about a steady state value, after 4 hours when using the sum of a 0.3 Hz square wave and a 0.75 hz square wave. This result indicates that the frequency content of the reference input is of primary importance for stability; not the saturating control input, unsymmetrical plant characteristics, or noisy derivative.

Fig. 11 shows the response to rapidly varying model parameters. (The effect on the negative definiteness of $\dot{V}$ is similar for varying model or plant parameters.) The model parameters were changed in approximately 0.5 sec. as follows: the gain from 26 to 22; the damping ratio from 0.25 to 0.03; and the natural frequency from 5.8 r/s to 7.1 r/s. Although appreciable control input saturation occurred, the system recovered in 3 sec. while maintaining its response within reasonable limits. Using ∇I resulted in poor recovery or overloading of the derivative circuit op-amps. Also, the system tracked a slowly varying model (i.e., an approx. 1%/sec. change in natural freq.) with tracking error similar to that for a constant model. Similar results were obtained for the other types of reference inputs. If the model parameters are varied quickly and continuously over a long enough time period, instability results. The length of this time period depends primarily on the size of the gains $1/\lambda_i$ on the integral terms. Large gains speed recovery but also shorten the time period over which the parameters may be varied without causing instability. The limiting factor here appears to be saturation of the control input. During rapid parameter variations heavy saturation of the control input occurs in the D/A which causes the integral terms to become erroneous, which in turn results in almost continuous saturation. This causes heavy saturation of the servo power amp which prevents the feedback necessary to the recovery of the integral terms and, even if parameter variation ceases, the system remains unstable.

Fig. 12 is the response to a disturbance input. The rather severe disturbance consists of holding the motor output shaft stationary for two sec.. Comments as to recovery are similar to those for varying parameters with the exception that here a 'minor' disturbance would be bearing down on the output shaft with sufficient force to reduce motor velocity by 25% while applying a constant plant input.

CONCLUSIONS

Although the saturation of the control input in the D/A and the noisy approximate derivative have a noticeably adverse influence on performance, the system is effectual and recovers readily from fairly severe parameter variations and disturbances. The system apparently adapts for saturation in the D/A by adjusting the period of saturation. This is not the case for saturation in the servo amps which causes a marked decrease in performance. The noise has a noticeable effect on the integral term (where the noise squared is integrated) of the adaptive gain associated with the derivative, Ik_2. If the noise level is high, such as when tracking a square wave, Ik_2 makes a negligible contribution to performance. Setting the gain on Ik_2 high enough may actually cause performance to deteriorate.

The modified design is very effective in overcoming the non-ideal aspects of the actual system. It's use greatly decreases tracking error, damps the error response and holds system operation within recoverable limits if for some reason, such as parameter variations or disturbances, the integral terms become erroneous.

The importance of the individual adaptive gains depends primarily on the type of reference input. Although all gains contribute to improved performance, performance does not decrease appreciably if the integral term associated with derivative, Ik_2, and the gain associated with the stiction friction, k_5, are deleted. Pk_2 makes a substantial contribution to improved performance.

Since the frequency content of the reference input, parameter variations, and disturbances affect the stability of the integral terms, the design could be improved considerably by including a criterion to bound the integral portion of the adaptive gains.

Acknowledgement

This research was partially supported by the National Aeronautics and Space Administration under grant NGL-22-010-018.

BIBLIOGRAPHY

[1] Price, C. F. and Koenigsberg, W. D., "Adaptive Control and Guidance for Tectical Missiles," Vol. 1 and 2, The Analytic Sciences Corporation, 6 Jacob Way, Reading, Massachusetts, June 1970.

[2] Monopoli, R. V., "Liapunov's Method for Adaptive Control System Design," IEEE Trans., AC-12, No. 3, pp. 334-335, June 1967.

[3] Lindorff, D. P., and Carroll, R. L., "Survey of Adaptive Control using Liapunov Design," Technical Report 72-7, Department of Electrical Engineering, University of Connecticut, Storrs, Connecticut, 1972.

[4] Parks, P. C., "Liapunov Redesign of Model-Reference Adaptive Control System," IEEE Trans., AC-11, No. 3, pp. 362-367, July 1966.

[5] Rang, E. R., "Adaptive Controllers Derived by Stability Consideration," Minneapolis-Honeywell Regulator Co., M.R. 7905, May 9, 1962.

[6] Monopoli, R. V. and Gilbart, J. W., "Model Reference Adaptive Control System with Feedback and Prefilter Adjustable Gain," Proc. of the Fourth Annual Princeton Conf. on Information Sciences and System, March 1970.

[7] Lasalle, J. P., "Asymptotic Stability Criteria," Proc. of Symposia in Applied Math., Amer. Math Soc., Providence, R.I., V. 13, 1962.

[8] Winsor, C. A. and Roy, R. J., "Design of Model Reference Adaptive Control System by Liapunov's Second Method," IEEE Trans., AC-13, No. 2, p. 204, April 1968.

[9] Gilbart, J. W., "Combination Open Loop, Closed Loop and Adaptive Compensation: It's Application to an Optical Tracking Telescope," Proceedings of the 1972 Int. Conference on Cybernetics and Society, Oct., Washington, D. C. (IEEE).

[10] Poter, B., and Tatnall, M. L., "Performance Characteristics of an Adaptive Hydralic-servo-mechanism," Int. J. Control, 1970, Vol. 11, No. 5, pp. 741-757.

[11] Gilbart, J. W. and Monopoli, R. V., "A Modified Liapunov Design for Model Reference Adaptive Control System," Proc. 7th Annual Allerton Conf. on Circuit and System Theory, pp. 32-33, October 1969.

[12] Graham, D. and McRuer, D., "Analysis of Nonlinear Control Systems," J. Wiley and Sons, Inc., p. 117, 1961.

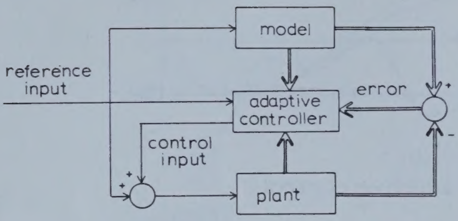

model
reference input
adaptive controller
error
control input
plant
+
-

Fig 1 general system

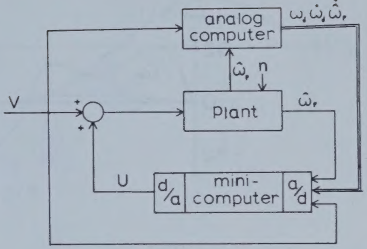

analog computer
$\omega_r\ \dot{\omega}_r\ \dot{\hat{\omega}}_r$
$\hat{\omega}_r$
n
V
Plant
$\hat{\omega}_r$
U
d/a
mini-computer
a/d

Fig 3 equipment

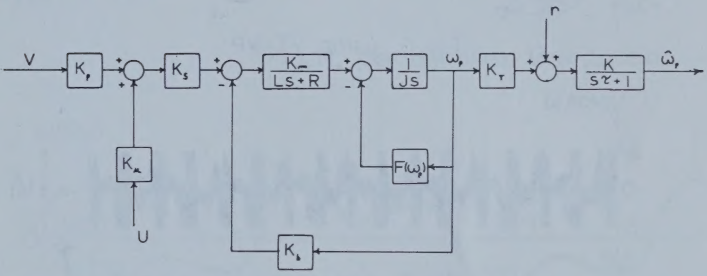

V
K_r
K_s
$\frac{K_m}{Ls+R}$
$\frac{1}{Js}$
ω_r
K_T
r
$\frac{K}{s\tau+1}$
$\hat{\omega}_r$
K_u
U
$F(\omega_r)$
K_b

Fig.2, plant

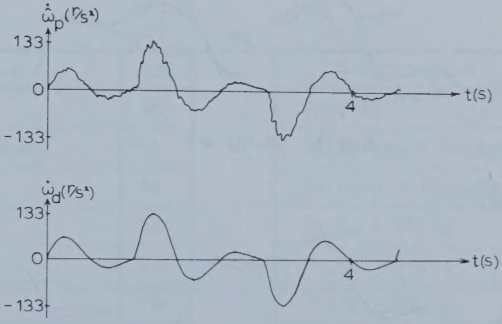


Fig. 4, derivatives-square wave

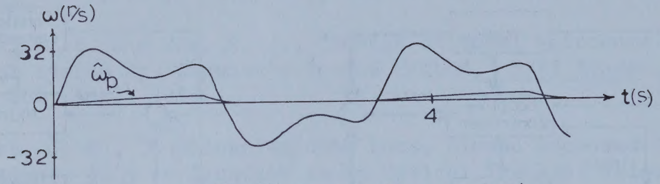


Fig.5, no adaptive control

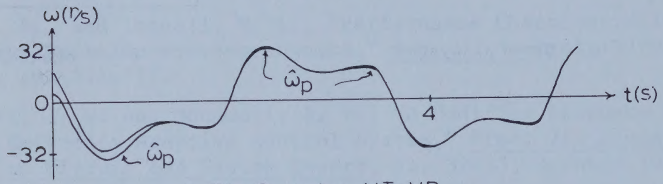


Fig.6, using ∀I+∀P

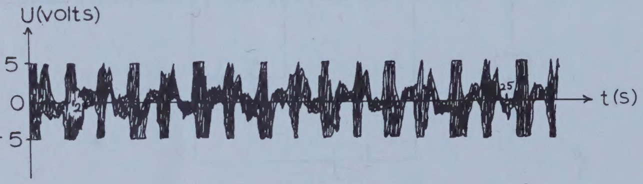


Fig.7, control input for fig 6

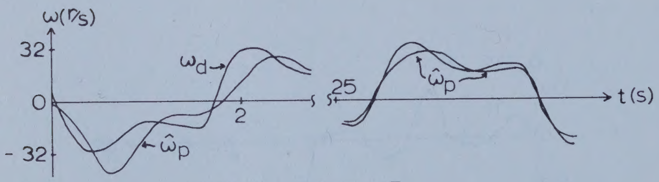


Fig.8, using ∀I

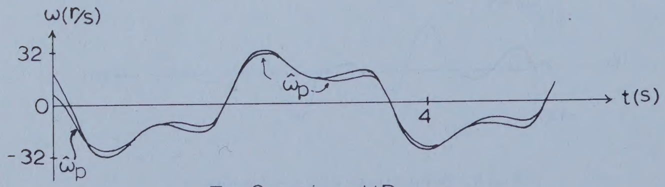


Fig.9, using ∀P

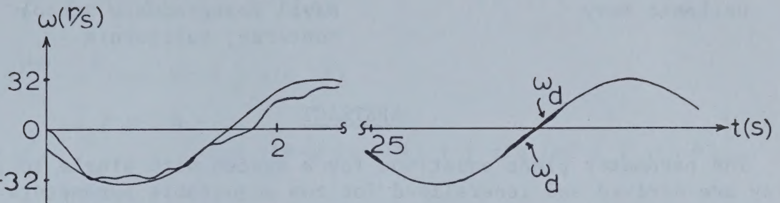


Fig.10, using ∀I

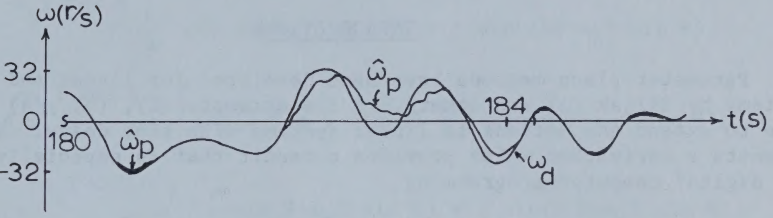


Fig.11, parameter variations

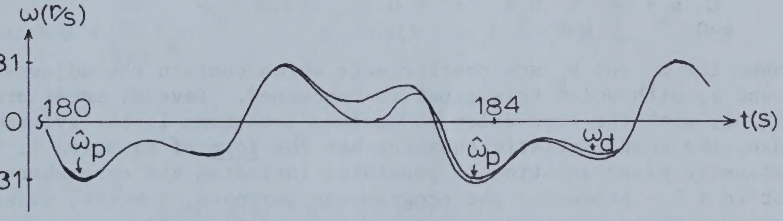


Fig.12, disturbance

Gains Used	IEM
∀I + ∀P	30
$Ik_1k_2k_3 + Pk_1k_2k_3$	31
$Ik_1k_3 + Pk_1k_3$	41
∀I	132
Ik_1k_3	132
∀P	71
$Pk_1k_3k_4k_5$	79

TABLE 1

Gains Used	IEM
∀I + ∀P	21
$IK_1k_2k_3k_4 + Pk_1k_2k_3k_4$	23
$Ik_1k_2k_3 + Pk_1k_2k_3$	120
$Ik_1k_4 + Pk_1k_4$	27
$Ik_2k_4 + Pk_2k_4$	28
$Ik_3k_4 + Pk_3k_4$	27
∀I	132
$Ik_1k_2k_3$	160
Ik_1k_3	160
∀P	34

PARAMETER PLANE METHODS FOR LINEAR SYSTEMS WITH TIME DELAY

A. HADJICONSTANDIS
Hellenic Navy

G. THALER
Naval Postgraduate School
Monterey, California

ABSTRACT

The parameter plane equations for a system with single lumped time delay are derived and generalized for two adjustable parameters arbitrarily located in the system. Practical interpretation of stability via Karmarkar's criterion is developed. The results are applied to compensation design for the roll mode of a vertical take-off aircraft.

INTRODUCTION

Parameter plane methods have been developed for linear and non-linear systems by Siljak (1) and others. A few attempts (2), (3), (4) have been made to extend the methods to linear systems with time delay. This paper presents a derivation which provides a result that is especially convenient for digital computer programming.

The characteristic equation of a system with time delay is of the form

$$\sum_{k=0}^{n} a_k s^k + \sum_{k=0}^{n} b_k e^{-sT} s^k = 0 \qquad (1)$$

where the a_k and b_k are coefficients which contain the adjustable parameters α and β, with which this paper is concerned. Several cases are shown in Fig. 1, with the time delay at various locations in the system. In each case, the characteristic equation has the form of equation 1, thus a general parameter plane solution is possible, including the case where terms exist for an α β - product. For programming purposes, however, it seems better to treat the α β - product case as a separate problem, and the solution is given in reference 5, but is not included here.

The general form of the characteristic equation to be considered is, therefore:

$$\sum_{k=0}^{n} d_k s^k + \sum_{k=0}^{n} e^{-Ts} f_k s^k + \alpha \sum_{k=0}^{n} e^{-Ts} b_k s^k + \beta \sum_{k=0}^{n} e^{-Ts} c_k s^k = 0 \qquad (2)$$

DERIVATION OF THE PARAMETER PLANE EQUATIONS

We are concerned, of course, with the locations of the roots of the characteristic equation as functions of the parameters α and β.

A. For Real Roots:

Let $s = -\sigma$, a real number. Then equation 2 becomes:

$$N_1 + N_2 + \alpha N_3 + \beta N_4 = 0 \qquad (3)$$

where N_1, N_2, N_3, N_4 are constants. Equation 3 is the equation of a straight line on the α, β parameter plane.

B. For Complex Roots:

$$\text{Let } s = \omega_n e^{j\theta} = \omega(\cos\theta + j\sin\theta)$$

$$s^k = \omega_n^k e^{jk\theta} = \omega^k(\cos k\theta + j\sin k\theta)$$

$$\text{then } e^{-sT}s^k = \omega_n^k e^{-\omega_n T\cos\theta}\left[\cos(k\theta - \omega_n T\sin\theta) + j\sin(k\theta - \omega_n T\sin\theta)\right]$$

$$\begin{aligned}
0 = {} & \sum_{k=0}^{n} d_k \omega_n^k (\cos k\theta + j\sin k\theta \\
& + \sum_{k=0}^{n} f_k e^{-\omega_n T\cos\theta}\omega_n^k\left[\cos(k\theta - \omega_n T\sin\theta) + j\sin(k\theta - \omega_n T\sin\theta)\right] \\
& + \alpha\sum_{k=0}^{n} b_k e^{-\omega_n T\cos\theta}\omega_n^k\left[\cos(k\theta - \omega_n T\sin\theta) + j\sin(k\theta - \omega_n T\sin\theta)\right] \\
& + \beta\sum_{k=0}^{n} c_k e^{-\omega_n T\cos\theta}\omega_n^k\left[\cos(k\theta - \omega_n T\sin\theta) + j\sin(k\theta - \omega_n T\sin\theta)\right]
\end{aligned}\tag{4}$$

Requiring reals and imaginaries to become zero independently:

$$\begin{aligned}
& \sum_{k=0}^{n} d_k \omega_n^k \cos k\theta + \sum_{k=0}^{n} f_k e^{-\omega_n T\cos\theta}\omega_n^k \cos(k\theta - \omega_n T\sin\theta) \\
& + \alpha\sum_{k=0}^{n} b_k e^{-\omega_n T\cos\theta}\omega_n^k \cos(k\theta - \omega_n T\sin\theta) \\
& + \beta\sum_{k=0}^{n} c_k e^{-\omega_n T\cos\theta}\omega_n^k \cos(k\theta - \omega_n T\sin\theta) = 0
\end{aligned}\tag{5A}$$

$$\begin{aligned}
& \sum_{k=0}^{n} d_k \omega_n^k \sin k\theta + \sum_{k=0}^{n} e^{-\omega_n T\cos\theta}\omega_n^k \sin(k\theta - \omega_n T\sin\theta) \\
& + \alpha\sum_{k=0}^{n} b_k e^{-\omega_n T\cos\theta}\omega_n^k \sin(k\theta - \omega_n T\sin\theta) \\
& + \beta\sum_{k=0}^{n} c_k e^{-\omega_n T\cos\theta}\omega_n^k \sin(k\theta - \omega_n T\sin\theta) = 0
\end{aligned}\tag{5B}$$

Note that these equations are of the form:

$$\begin{aligned}
D_1 + \alpha B_1 + \beta C_1 &= 0 \\
D_2 + \alpha B_2 + \beta C_2 &= 0
\end{aligned}\tag{6}$$

and for any chosen value of s the B's, C's and D's are constants so a simultaneous solution for α and β is available.

For ease in computerizing, we find recursion relationships to replace trigonometric functions.

$$\text{Let } Y_k = \sin(k\theta - \omega_n T \sin\theta) \tag{7}$$

$$X_k = \cos(k\theta - \omega_n T \sin\theta)$$

noting that $\cos\theta = \zeta$, $\sin\theta = \sqrt{1-\zeta^2}$, and manipulating:

$$X_{-1} = -\left(\sqrt{1-\zeta^2}\,\sin\left[\omega_n T\sqrt{1-\zeta^2}\right] + \zeta\cos\left[\omega_n T\sqrt{1-\zeta^2}\right]\right)$$

$$X_0 = \cos\omega_n T\sqrt{1-\zeta^2} \tag{8}$$

$$X_1 = \sqrt{1-\zeta^2}\,\sin\left[\omega_n T\sqrt{1-\zeta^2}\right] - \zeta\cos\left[\omega_n T\sqrt{1-\zeta^2}\right]$$

By induction applied to equation 7

$$X_{k+1} = -2\zeta X_k - X_{k-1}$$

$$Y_{k+1} = -2\zeta Y_k - Y_{k-1} \tag{9}$$

$$Y_k = \frac{1}{\sqrt{1-\zeta^2}}\left[\zeta X_k + X_{k-1}\right]$$

Let us also define:

$$U_k = \sin k\theta$$

$$T_k = \cos k\theta \tag{10}$$

from which

$$T_{-1} = -\zeta$$

$$T_0 = 1 \tag{11}$$

$$T_1 = -\zeta$$

By induction:

$$T_{k+1} = -2\zeta T_k - T_{k-1}$$

$$U_k+1 = -2\zeta U_k - U_{k-1}$$

$$U_k = \frac{1}{\sqrt{1-\zeta^2}}(\zeta T_k + T_{k-1})$$

Equations 5 thus become (note that $-\omega_n T\cos\theta = \zeta\omega_n T$)

$$\sum_{k=0}^{n} d_k\omega_n^k T_k + \sum_{k=0}^{n} f_k e^{\zeta\omega_n T}\omega_n^k X_k + \alpha\sum_{k=0}^{n} b_k e^{\zeta\omega_n T}\omega_n^k X_k$$

$$+\beta\sum_{k=0}^{n} c_k e^{\zeta\omega_n T}\omega_n^k X_k = 0$$

$$\sum_{k=0}^{n} d_k\omega_n^k(\zeta T_k + T_{k-1}) + \sum_{k=0}^{n} f_k e^{\zeta\omega_n T}\omega_n^k(\zeta X_k + X_{k-1}) \tag{13}$$

$$+\alpha\sum_{k=0}^{n} b_k e^{\zeta\omega_n T}\omega_n^k(\zeta X_k + X_{k-1}) + \beta\sum_{k=0}^{n} c_k e^{\zeta\omega_n T}\omega_n^k(\zeta X_k + X_{k-1}) = 0$$

Equation 13 is the desired result.

ROOT LOCUS STUDY OF A VERTICAL TAKE-OFF AIRCRAFT

Fig. 2 is the block diagram for the roll-mode of a vertical take-off aircraft. The time delay is variable and may be within the limits $.15 \le T \le .30$. We want a stable loop with dominant roots in the range of pilot preference, i.e., $\zeta = .7$, $1 \le \omega_n \le 3$. It is proposed that this be accomplished by using velocity feedback and adjusting the parameters K_1 and K_2.

By inspection of the transfer function the feedback factor which may be written $K_1\ (s + K_2/K_1)$, provides a "free" zero and also loop gain adjustment. Thus a root locus analysis is feasible if we neglect the time delay. Fig. 3 shows the root locus diagram for one possible location of the free zero which permits acceptable roots. Note that the loop can be stabilized and acceptable root locations obtained with many possible locations of the free zero. It is even possible to put the zero in the right half plane if positive feedback is used. The effect of the time delay will deteriorate the system behavior, but quantitative design using root locus is inconvenient. We, therefore, use the preliminary results of Fig. 3 as a guide for parameter plane studies. Fig. 3 shows that there will be three pairs of complex roots. The desired pair, a pair near the control surface resonance ($\omega_n \cong .7$), and a very remote pair associated with the servo poles. It also indicates that with small time delay, the required gains may be near the values $K_1 = 100$, $K_2 = 200$.

STABILITY ANALYSIS IN THE PARAMETER PLANE

Use of the parameter plane program based on equation 13 provides the curves of Fig. 4 (for $T = .30$). The curves are used as usual -- pick a desired root value, locate the root point on the curve, read off its coordinates (α, β), calculate the corresponding K_1, K_2. All other roots of the characteristic equation are defined by this choice and can be read from the curves if the curves have been completed in sufficient detail. To avoid misinterpretation, we wish to define stability limits on the parameter plane before choosing an operating point. The purpose of the stability limits is to point out an area on the parameter plane such that choice of an operating point within this area guarantees that all roots are in the left half plane.

Karmarkar's (6) stability criterion states:

"Consider CE(s) to be a transcendental polynomial

$$CE(s) = \sum_{p,v}^{m,n} a_{pr} s^p e^{rs} \tag{14}$$

where a_{pr} = real coefficients ($a_{mn} \ne 0$), p, n are integers, and

$$\sum_{p,r}^{m,n} a_{pr} s^p e^{rs} = \begin{bmatrix} s^0 & s^1 & \cdots & s^n \end{bmatrix} \begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0n} \\ a_{10} & a_{11} & \cdots & a_{1n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m0} & a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} e^{0s} \\ e^{r_1 s} \\ \cdot \\ \cdot \\ \cdot \\ e^{r_n a} \end{bmatrix} \tag{15}$$

For CE(s) to be stable, it is necessary and sufficient to find an integer k such that, starting with the line $\beta = \beta_0$ on the α-β plane, the total number, λ, of alternating intersections of the $\zeta = 0$ curve with the $\alpha = \alpha_0$ and $\beta = \beta_0$ lines is

$$\lambda = 4kn+m \tag{16}$$

when ω ranges over the interval

$$0 \le \omega \le 2\pi k + \epsilon/2, \text{ for } 0 \le \epsilon/2 \le \pi/2, \text{ or}$$
$$\text{for } \epsilon = \pi\delta \quad 0 \le \omega \le 2\pi(k + \delta/4), \text{ for } 0 \le \delta \le 1 \tag{17}$$

To illustrate the meaning of this criterion, consider the sketch of Fig. 5. We consider only the $\zeta = 0$ curve on the α-β plane, choose an operating point $\alpha = \alpha_0$ and $\beta = \beta_0$, drawing the auxiliary lines α_0 and β_0 for convenience. Starting on the $\zeta = 0$ curve at $\omega = 0$ and moving in direction of increasing ω we alternately cross the β_0 and α_0 lines. We consider both the number of intersections and the frequency range in which the intersections occur. Fig. 5 was prepared for the equation CE(s) = $(s^3 + 2s^2 + s)e^s + as + b = 0$, for which m=3 and n=1. Using equations 16 and 17 and trying k = 2, $\lambda = 4 \cdot 2 \cdot 1 + 3 = 11$ and the frequency range is $0 \le \omega \le 14.152$. On Fig. 5 we count 11 intersections and the frequency at intersection 11 is less than 14.15, so the system is stable for operating point at $\alpha = \alpha_0$ and $\beta = \beta_0$. Obviously it is also stable for any other operating point which maintains these conditions, and on Fig. 5 this is clearly the area enclosed by the 1-2-3 loop and the real axis.

Applying this stability criterion to the aircraft, on Fig. 4 are superimposed the stability boundaries for this system with T = 0.3. To assess the overall effect of the time delay, we must compute separate parameter planes for each value of T and observe the change in the stability boundary and in the location of the operating point required to maintain desired values of the complex roots. These results are shown in Fig. 6 and in Table I.

Inspection of Fig. 4 and comparison with the root locus of Fig. 3 raises a question: The root locus indicates that in addition to the desired root at $\zeta \cong .8$, $\omega_n \cong 1.6$, there will be complex roots at $\zeta \cong .3$, $\omega_n \cong .7$; why does not Fig. 4 predict such roots? The answer is given on Fig. 7: constant ζ and constant ω_n curves for the narrow range including the structural resonance cover the entire plane, but the computer must be programmed to scan this range with a very fine increment or the sub-family of curves may be skipped.

CONCLUSION

Systems with a single lumped time delay can be analyzed using parameter plane methods. A computer program has been developed to provide the necessary curves, and these curves have been used successfully in design studies.

REFERENCES

1. Siljak, D. D. "Nonlinear Systems" John Wiley and Sons, Inc. 1969.

2. Eisenberg, L. "Design of Feedback Control Systems with Transport Lag by Parameter Plane Techniques" Doctoral Dissertation Newark College of Engineering 1966

3. Lamberski, S. E. "Parameter Plane Solution of Linear Feedback Control Systems with Alpha-Beta Product and Transport Lag" M.S. Thesis, Newark College of Engineering 1967.

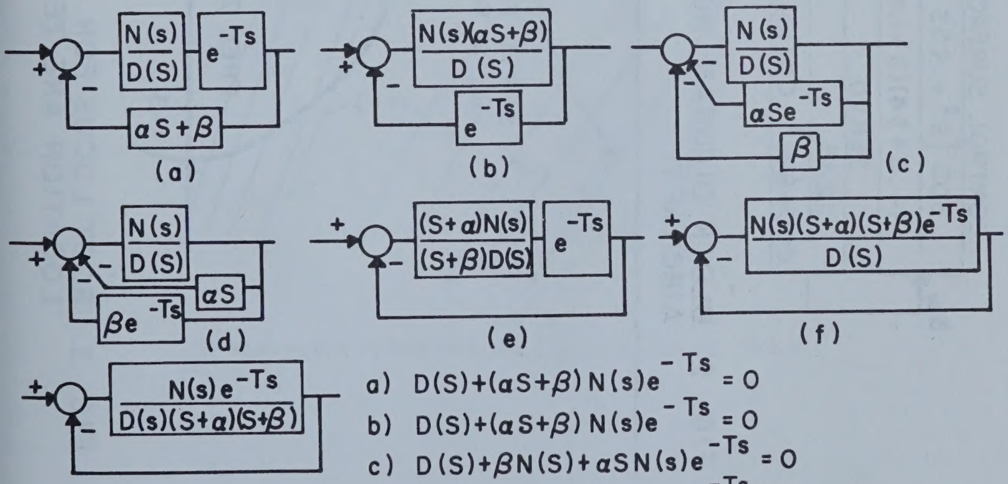
4. Karmarkar, J. S. "Stability Analysis of Systems with Time Delay" Proc. IEEE v 117, No. 7, July 1970.

5. Hadjiconstandis, A. C. "Parameter Plane Methods for Linear Systems with Time Delay" M.S. Thesis, Naval Postgraduate School March 1973.

TABLE I

VARIATION OF "α" AND "β" VERSUS "T" $(\omega_n,\zeta) = (1.6,.8)$

T sec	α units	β units	K_p	K_ϕ
0.0	7340	14680	95	190
0.5	2853	2962	37	38.4
0.1	5722	10568	74.1	136.6
0.15	5574	9453	72.2	122.5
0.2	5280	8279.4	68.4	107.2
0.25	4893	7130	63.3	92.3
0.3	4463.6	6064.1	57.8	78.6
0.35	4026.4	5110.3	52.2	67.5



a) $D(S)+(\alpha S+\beta)N(s)e^{-Ts}=0$

b) $D(S)+(\alpha S+\beta)N(s)e^{-Ts}=0$

c) $D(S)+\beta N(S)+\alpha SN(s)e^{-Ts}=0$

d) $D(S)+\alpha SN(s)+\beta N(s)e^{-Ts}=0$

e) $(S+\beta)D(S)+(S+\alpha)N(s)e^{-Ts}=0$

f) $D(S)+N(s)(S+\alpha)(S+\beta)e^{-Ts}=0$

g) $D(S)(S+\alpha)(S+\beta)+N(s)e^{-Ts}=0$

IG. I. BLOCK DIAGRAMS OF SYSTEMS WITH TIME DELAY.

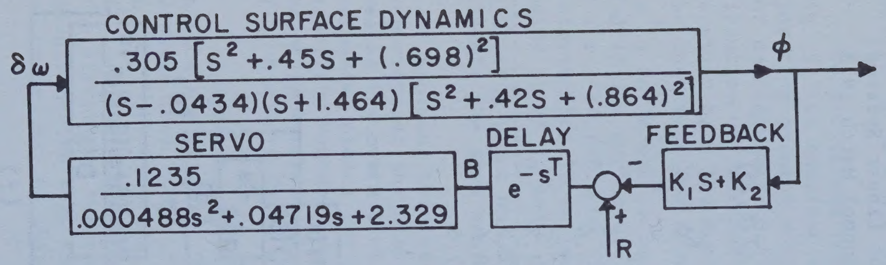

CONTROL SURFACE DYNAMICS
$\frac{.305\left[S^2+.45S+(.698)^2\right]}{(S-.0434)(S+1.464)\left[S^2+.42S+(.864)^2\right]}$
$\delta\omega$
ϕ
SERVO
$\frac{.1235}{.000488s^2+.04719s+2.329}$
B
DELAY
e^{-sT}
FEEDBACK
K_1S+K_2
-
+
R

FIG. 2. BLOCK DIAGRAM, ROLL MODE OF A VERTICAL TAKEOFF AIRCRAFT.

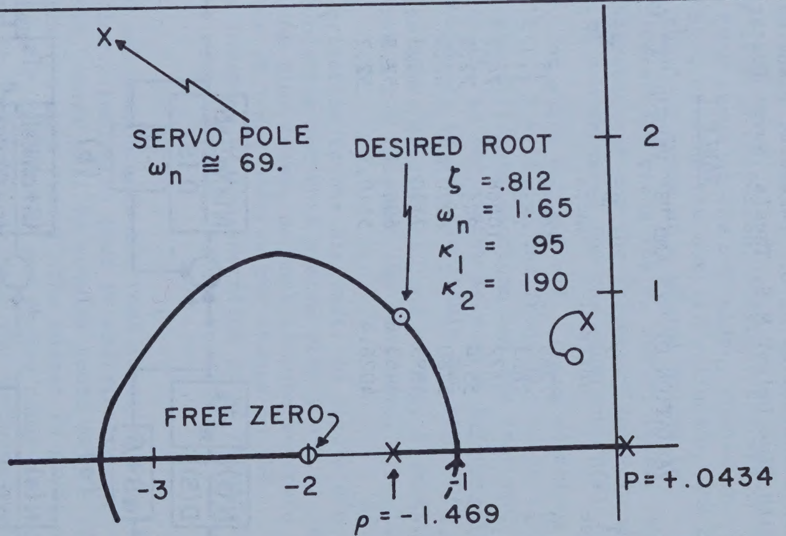


FIG. 3. ROOT LOCUS FOR ROLL MODE. T=0 ACHOSEN LOCATION ARE ZERO.

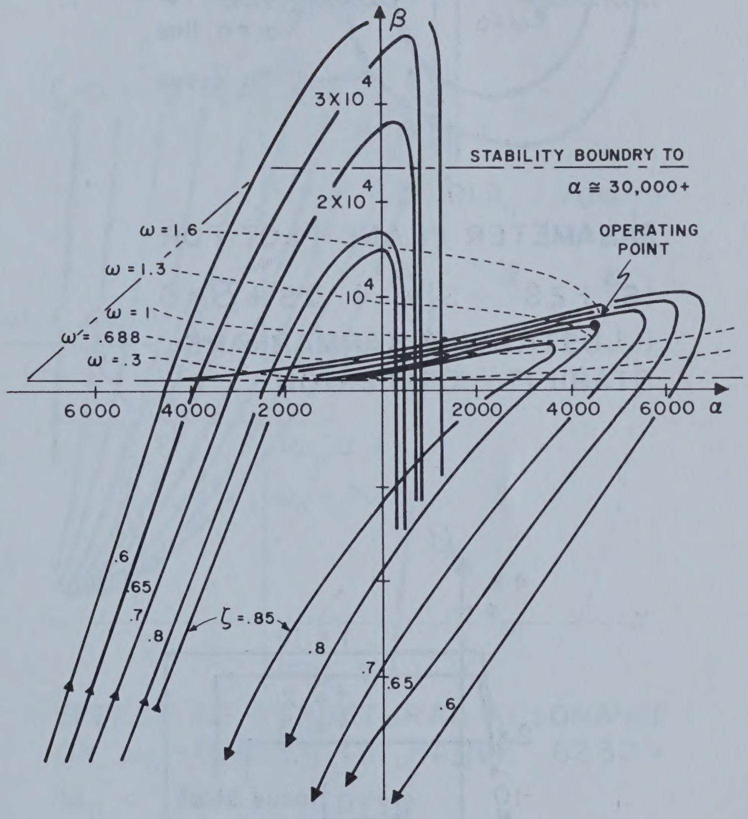


FIG. 4.
PARAMETER PLANE FOR ROLL LOOP OF FIG. 2. T = 0.3

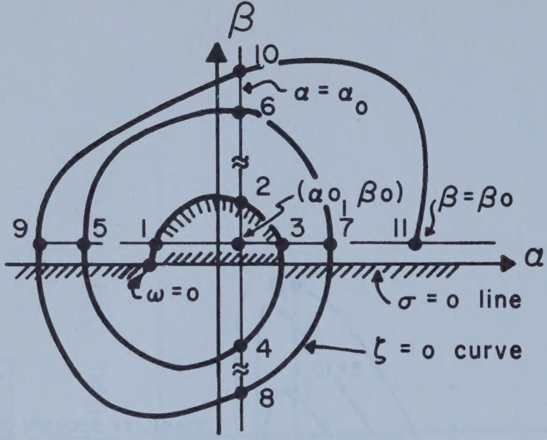

β
10
α = α₀
6
2
(αo, βo)
β=βo
9
5
1
3
7
11
α
ω=o
σ = o line
4
ζ = o curve
8

FIG. 5.
PARAMETER PLANE PLOT FOR
$(S^3 + 2S^2 + S)e^S + \alpha S + \beta = 0$
ILLUSTRATING KARMARKARS
STABILITY CRITERION

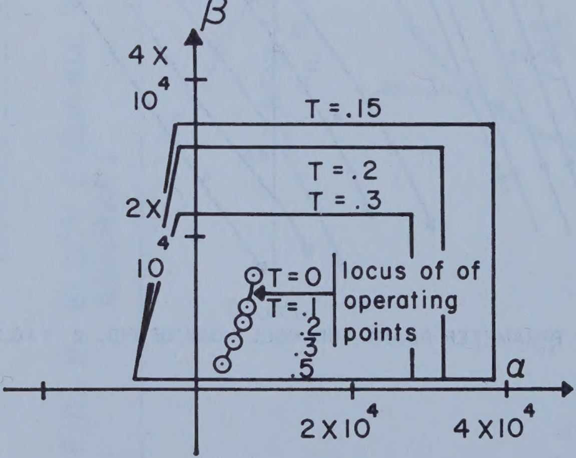


FIG. 6.
EFFECT OF TIME DELAY ON MAXIMUM ROOT REGIONS AND ON THE OPERATING POINT

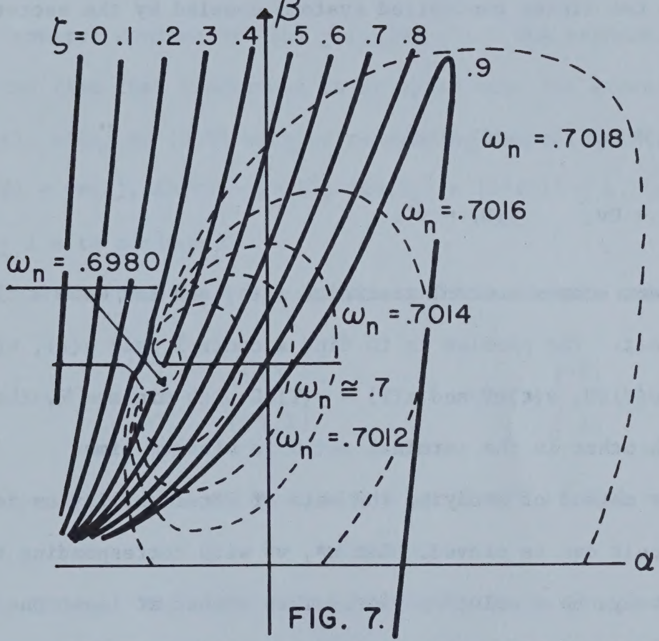


FIG. 7.

EFFECT OF STRUCTURAL RESONANCE ON ω_n -CURVES (IN RANGE $.6280 < \omega_n < .7018$)

ON RENDEZVOUS WITH TERMINAL CONSTRAINT*

DONG HAK CHYUNG
Dept. of Electrical Engineering, Univ. of Iowa, Iowa City, Iowa 52242

ABSTRACT

Two linear systems are required to rendezvous in a given terminal area in minimum time.

We consider two linear controlled systems modeled by the vector differential equations

$$\dot{x} = Ax + Bu, \qquad x(0) = x^\circ$$

$$\dot{y} = Cy + Dv, \qquad y(0) = y^\circ.$$

Let U, V be compact convex control restraint sets, and let G be a closed convex terminal set. The problem is to find a control pair $u(t)$, $v(t)$ on $[0,T]$ such that $u(t)\varepsilon U$, $v(t)\varepsilon V$ and $x(T) = y(T)\varepsilon G$ with minimum T, that is, x and y meet each other in the terminal set G in minimum time.

By a similar method of studying the sets of attainability as in Ref. 1, the following result can be proved. Let u^*, v^* with corresponding responses x^*, y^*, respectively, be a solution pair. Then either at least one of the controls u^*, v^* is a time optimal control to the target G or there exist nontrivial solutions $p(t)$, $q(t)$ of the differential equations

$$\dot{p} = -pA, \qquad \dot{q} = -qC$$

such that

$$P(t)Bu^*(t) = \max p(t)Bu, \qquad q(t)Dv^*(t) = \max q(t)Dv$$

for all $u\varepsilon U$, $v\varepsilon V$ for almost all t in $[0,T]$. In this case, if $x(T) = y(T)$

---

*This research was supported by NSF Grant GK-31628.

is not a boundary point of the target set G, then u*, v* are time optimal rendezvous controls for x and y with no terminal target set constraint.

As an example consider the minimum time rendezvous problem for two linear scalar systems

$$\ddot{x} = u, \qquad x(0) = 0, \qquad \dot{x}(0) = -2\sqrt{2}+1$$

$$\ddot{y} = v, \qquad y(0) = 4(\sqrt{2}-1), \qquad \dot{y}(0) = -2\sqrt{2}+1$$

with the control constraints $|u| \leq 1$, $|v| \leq 1$. The terminal constraint is that at the time they rendezvous their speed must not exceed 1. That is, find $u^*(t)$, $v^*(t)$ on $[0,T]$ with corresponding responses $x^*(t)$, $y^*(t)$ such that $x^*(T) = y^*(T)$, $\dot{x}^*(T) = \dot{y}^*(T)$, $|\dot{x}^*(T)| = |\dot{y}^*(T)| \leq 1$, $|u^*(t)| \leq 1$, $|v^*(t)| \leq 1$ with minimum T.

The minimum time is T = 2 and the optimal controls are given by

$$u^*(t) = \begin{cases} 1, & [0,\sqrt{2}] \\ -1, & [\sqrt{2},2], \end{cases} \qquad v^*(t) = \begin{cases} -1, & [0,2\sqrt{2}-2] \\ 1, & [2\sqrt{2}-2,2]. \end{cases}$$

The rendezvous point is $x^*(2) = y^*(2) = -2$, $\dot{x}^*(2) = \dot{y}^*(2) = -1$.

REFERENCE

1. D. H. Chyung, "Time Optimal Rendezvous of Three Linear Systems," J. of Optimization Theory and Applications, Vol. 12, September 1973.

APPLICATION OF THE EPSILON METHOD TO OPTIMAL CONTROL COMPUTATION

S. MUKHERJEE and N. K. SINHA
Group on Simulation, Optimization & Control
Department of Electrical Engineering
McMaster University, Hamilton, Ontario, Canada

ABSTRACT

An efficient gradient method for optimization, proposed recently by Fletcher, is applied to make the epsilon method more useful for solving optimal control problems. A number of examples are used for comparing the results obtained using earlier gradient techniques.

SUMMARY

Balakrishnan's epsilon method [1,2] presents a different and computationally attractive method for solving optimal control problems since it does not require the solution of differential equations. Results on the computational implementation of the method have been published by some authors[3,4]. Whereas their results are promising and show the method to be successful, better results could have been obtained if they had used more efficient gradient methods rather than the less efficient gradient method and the Newton-Raphson method. In particular the method proposed by Fletcher[5] has been shown to be much faster than the older methods[6], when applied to algebraically constrained minimization problems. It cannot in general be applied to the general optimal control problem due to the dimensionality introduced by differential constraints.

It has been reported[3,4] that the Newton-Raphson method gives better results than those obtained by using gradient methods they used. In the present investigation it has been found that Fletcher's method yield almost equivalent results as those obtained by Newton-Raphson method.

It is therefore felt that with the availability of efficient gradient methods epsilon technique of computing optimal control can be more useful in its application to a wider class of problems.

REFERENCES

1. A.V. Balakrishnan: "On a new computing technique in optimal control", SIAM Jour. Control, May 1968, pp. 149-173.
2. A.V. Balakrishnan: "On a new computing technique in optimal control and its application to minimal time flight profile optimization", J. Optimization Theory & Applications, July, 1969.
3. J.M. Taylor and C.T. Constantinides: "Optimization: Application of the Epsilon Method", IEEE Trans. vol. AC-17, Feb. 1972, pp. 128-131.
4. L.J. Taylor, Jr., J.H. Smith, and K.W..Illiff, "Experience using Balakrishnan's epsilon technique to compute optimum flight profiles" Presented at the AIAA 7th Aerospace Sciences Meeting, New York, N.Y. Jan. 20-22, 1969.
5. R. Fletcher: "A new approach to variable metric algorithms", Computer Journal, vol. 13, Aug. 1970, pp. 317-322.
6. J.W. Bandler, N.D. Markettos and N.K.Sinha: "Optimum system modelling using recent gradient methods", International Journal of System Science, vol. 4, No. 1, January 1973, pp. 33-44.

DISCRETE OPTIMIZATION

CHRISTAKIS CHARALAMBOUS
Department of Combinatorics and Optimization
University of Waterloo
Waterloo, Ontario, Canada.

Abstract. This paper presents some theoretical ideas for the discrete nonlinear programming problem. The main objective is to construct a hypercube inside of which the discrete optimum will lie. It is shown that for quadratic functions this hypercube can be constructed by knowing the inverse of the hessian matrix.

INTRODUCTION

Discrete optimization techniques are of great interest to engineers and applied mathematicians. The engineer has a practical or a semipractical problem demanding solution, while the applied mathematician is challenged primarily, by the difficult task of obtaining theoretical conditions for optimality. Over the past decade the applied mathematicians proposed a number of algorithms to solve the linear integer programming problem [1],[2], but very little has been done in the area of nonlinear integer programming. Over the past few years the engineers made an attempt to give a solution to the problem but they did not have much success [3], [4]. Their methods are "heuristic" and they do not have much mathematical structure in them. In this paper an attempt is made to give a mathematical structure to solve the nonlinear integer programming problem. The main objective is to find upper and lower bounds on the parameters, within which the integer solution will lie. It is shown that for quadratic functions these bounds can be obtained by knowing the inverse of the hessian matrix.

THEORY

The nonlinear integer programming problem can be stated as:

$$\text{minimize} \quad F(\underset{\sim}{x}) \tag{1}$$

Subject to

$$g_j(\underset{\sim}{x}) \geq 0, \; j = 1, 2, \ldots, m \tag{2}$$

where

$$\underset{\sim}{x} \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{3}$$

is an n-dimensional column vector containing the adjustable parameters and

$$x_i = \text{Integer},\ i = 1, 2, \ldots, n \tag{4}$$

$F(\underset{\sim}{x}), g_1(\underset{\sim}{x}), g_2(\underset{\sim}{x}), \ldots, g_m(\underset{\sim}{x})$ are, in general, nonlinear functions of the parameters.

Unconstrained Optimization. Let us first consider the case where we have the only constraints that the parameters are integers. Assume that we start from an integer point $\bar{x}^0$. Then any integer point which is at least as good as the one which we started must lie inside or on the boundary of the region R, where

$$R \triangleq \{\underset{\sim}{x} | F(\underset{\sim}{x}) - F(\bar{\underset{\sim}{x}}^0) \le 0\} \tag{5}$$

In other words, if we denote the integer optimum by $\bar{\underset{\sim}{x}}$, then $\bar{\underset{\sim}{x}} \in R$. Let us further define

$$\left.\begin{aligned} x_{\ell i} &\triangleq \min x_i \\ x_{ui} &\triangleq \max x_i \end{aligned}\right\} \underset{\sim}{x} \in R,\ i = 1, 2, \ldots, n \tag{6}$$

then, $\bar{\underset{\sim}{x}} \in R_1$ where

$$R_1 \triangleq \{\underset{\sim}{x} | \underset{\sim}{x}^\ell \le \underset{\sim}{x} \le \underset{\sim}{x}^u\} \tag{7}$$

$$\underset{\sim}{x}^\ell \triangleq \begin{bmatrix} x_{\ell 1} \\ x_{\ell 2} \\ \cdot \\ \cdot \\ \cdot \\ x_{\ell n} \end{bmatrix}, \qquad \underset{\sim}{x}^u \triangleq \begin{bmatrix} x_{u1} \\ x_{u2} \\ \cdot \\ \cdot \\ \cdot \\ x_{un} \end{bmatrix} \tag{8}$$

$x_{\ell i}$ and x_{ui} are the lower and upper bounds of the ith component of the vector $\underset{\sim}{x}$. For example see Figure 1. $\overset{\vee}{\underset{\sim}{x}}$ denotes the continuous optimum. Obviously we have to assume that the set $S \triangleq \{x | F(x) \le c\}$ is bounded for every value of $c \le F(\bar{\underset{\sim}{x}}^0)$. This ensures that any local minimum is located at a finite point. Let $\bar{x}_{\ell i}$ and $\bar{x}_{ui}$ denote the closest integer points for $x_{\ell i}$ and x_{ui} for x_i increasing and x_i decreasing respectively. Then $\bar{\underset{\sim}{x}} \in \bar{R}_1$, where

$$\bar{R}_1 \triangleq \{\underset{\sim}{x} | \bar{\underset{\sim}{x}}^\ell \le \underset{\sim}{x} \le \bar{x}^u\} \tag{9}$$

For example, in Figure 1 the region $\bar{R}_1$ is enclosed by the shaded area. Therefore we can construct a hypercube inside of which the integer minimum will lie.

Assuming that there is only one minimum and one maximum of (6) these can be found by using the Lagrange function,

$$V_i(\underset{\sim}{x}, \lambda) = x_i + \lambda(F(\underset{\sim}{x}) - F(\bar{\underset{\sim}{x}}^0)),\ i = 1, 2, \ldots, n \tag{10}$$

where λ is the Lagrange multiplier. If we further assume that the gradients of $F(\underset{\sim}{x})$ with respect to $\underset{\sim}{x}$ exist, the following equations must hold at the optimum point.

$$\underset{\sim}{e}_i + \lambda \underset{\sim}{\nabla} F(\underset{\sim}{x}) = \underset{\sim}{0} \tag{11}$$

and

$$F(\underset{\sim}{x}) = F(\bar{\underset{\sim}{x}}^0) \tag{12}$$

where $\underset{\sim}{e}_i$ is an n-dimensional unit vector having unity at the ith element, and

$$\underset{\sim}{\nabla} \triangleq [\partial F/\partial x_1 \ldots \partial F/\partial x_n]^T \tag{13}$$

Equations (11) and (12) represent, in general n nonlinear equations to be solved. Equations (11) and (12) suggest that the original problem (6) can be solved by minimizing the following unconstrained objective function

$$U_i(\underset{\sim}{x}) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \left[\frac{\partial F(\underset{\sim}{x})}{\partial x_j}\right]^2 + [F(\underset{\sim}{x}) - F(\bar{\underset{\sim}{x}}^0)]^2 , \quad i = 1, 2, \ldots, n \tag{14}$$

whose minimum is zero.

Although the starting integer point $\bar{\underset{\sim}{x}}^0$ can in principal be anywhere, to be able to cut down the region $\bar{R}_1$ we can do the following. Find the continuous optimum $\check{\underset{\sim}{x}}$ and if its components are integer stop; otherwise calculate the value of the function at the 2^n integer points surrounding $\check{\underset{\sim}{x}}$ and take as $\bar{\underset{\sim}{x}}^0$ the integer point with the smallest objective function.

For minimax objectives the gradients are discontinuous at the points where two or more maxima are equal and we cannot use (11) directly. To overcome this difficulty we can use the generalized least pth objective function of Bandler and Charalambous with sufficiently large values of p [5].

Quadratic Objective Function. Let us assume that we have a quadratic function of the form

$$F(\underset{\sim}{x}) = \frac{1}{2}(\underset{\sim}{x} - \check{\underset{\sim}{x}})^T \underset{\sim}{A} (\underset{\sim}{x} - \check{\underset{\sim}{x}}) + \check{F} \tag{15}$$

where $\check{F}$ denotes the optimum value of $F(\underset{\sim}{x})$ and $\underset{\sim}{A}$ is an n×n dimensional positive definite symmetric matrix.

Theorem. For the quadratic function shown in (15) the lower and upper bounds on the parameters are given by

$$\left.\begin{aligned} x_{\ell i} &= \check{x}_i - \Delta x_i \\ x_{ui} &= \check{x}_i + \Delta x_i \end{aligned}\right\} \quad i = 1, 2, \ldots, n \tag{16}$$

where

$$\Delta x_i = \left(\frac{2(F(\bar{\mathbf{x}}^0) - \check{F})}{a_{ii} - (\mathbf{a}^i)^T \mathbf{A}_i^{-1} \mathbf{a}^i}\right)^{\frac{1}{2}} \tag{17}$$

$\mathbf{A}_i$ is an $(n - 1) \times (n - 1)$ dimensional matrix which is obtained from $\mathbf{A}$ by deleting the ith row and the ith column and $\mathbf{a}^i$ is the ith column of $\mathbf{A}$ with the ith element deleted.

Proof. The gradient vector of (15) is given by

$$\nabla F(\mathbf{x}) = \mathbf{A}(\mathbf{x} - \check{\mathbf{x}}) \tag{18}$$

From (11)

$$\mathbf{A}_i(\mathbf{x}^i - \check{\mathbf{x}}^i) = -(x_i - \check{x}_i)\mathbf{a}^i$$

where $\mathbf{x}^i$ is an $(n - 1)$ dimensional column vector which is obtained from $\mathbf{x}$ by deleting the ith element. Therefore,

$$\mathbf{x}^i - \check{\mathbf{x}}^i = -(x_i - \check{x}_i)\mathbf{A}_i^{-1}\mathbf{a}^i \tag{19}$$

Let $\mathbf{e}_{1i}$ and $\mathbf{e}_{2i}$ are $(i - 1)$ and $(n - i)$ column vectors respectively. Then,

$$\left[\mathbf{e}_{1i}^T \,\vdots\, 1 \,\vdots\, \mathbf{e}_{2i}^T\right] \mathbf{A} \begin{bmatrix} \mathbf{e}_{1i} \\ \hline 1 \\ \hline \mathbf{e}_{2i} \end{bmatrix} = \left[\mathbf{e}_{1i}^T \,\vdots\, 1 \,\vdots\, \mathbf{e}_{2i}^T\right] \begin{array}{c} \\ (i) \\ \\ \end{array} \begin{array}{c} \quad\quad\; (i) \\ \left[\begin{array}{c|c|c} \mathbf{A}_{1i} & \mathbf{a}_{1i} & \mathbf{A}_{2i} \\ \hline \mathbf{a}_{1i}^T & a_{ii} & \mathbf{a}_{2i}^T \\ \hline \mathbf{A}_{3i} & \mathbf{a}_{2i} & \mathbf{A}_{4i} \end{array}\right] \end{array} \begin{bmatrix} \mathbf{e}_{1i} \\ \hline 1 \\ \hline \mathbf{e}_{2i} \end{bmatrix}$$

$$= (\mathbf{e}_{1i}^T\mathbf{A}_{1i}\mathbf{e}_{1i} + \mathbf{e}_{1i}^T\mathbf{A}_{2i}\mathbf{e}_{2i} + \mathbf{e}_{2i}^T\mathbf{A}_{3i}\mathbf{e}_{1i} + \mathbf{e}_{2i}^T\mathbf{A}_{4i}\mathbf{e}_{2i})$$

$$+ 2(\mathbf{e}_{1i}^T\mathbf{a}_{1i} + \mathbf{e}_{2i}^T\mathbf{a}_{2i}) + a_{ii}$$

$$= \mathbf{e}^{i^T}\mathbf{A}_i\mathbf{e}^i + 2\mathbf{e}^{i^T}\mathbf{a}^i + a_{ii} > 0 \tag{20}$$

where $\mathbf{e}^i = [\mathbf{e}_{1i}^T \,\vdots\, \mathbf{e}_{2i}^T]^T$. The inequality is due to the fact that $\mathbf{A}$ is positive definite.

Let $\mathbf{e}^i = -\mathbf{A}_i^{-1}\mathbf{a}^i$ then substituting (19) into (12) and using (15) and (20) we have

$$\tfrac{1}{2}(x_i - \check{x}_i)^2 \left[\mathbf{a}^{i^T}\mathbf{A}_i^{-1}\mathbf{A}_i\mathbf{A}_i^{-1}\mathbf{a}^i - 2\mathbf{a}^{i^T}\mathbf{A}_i^{-1}\mathbf{a}^i + a_{ii}\right] = F(\bar{\mathbf{x}}^0) - \check{F}$$

or

$$x_i = \check{x}_i \pm \left(\frac{2(F(\mathbf{x}^0) - \check{F})}{a_{ii} - \mathbf{a}^{iT}\mathbf{A}_i^{-1}\mathbf{a}^i}\right)^{\frac{1}{2}} \tag{21}$$

By using the fact that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ it can be easily proved that

$$h_{ii} = \frac{1}{a_{ii} - \underset{\sim}{a}^{i^T} \underset{\sim}{A}_i^{-1} \underset{\sim}{a}^i} \tag{22}$$

where h_{ii} is the ith diagonal element of the inverse of $\underset{\sim}{A}$. (Equation (22) was suggested by Burns and Carter). Therefore,

$$\Delta x_i = \left(2h_{ii}(F(\bar{\underset{\sim}{x}}^0) - \check{F})\right)^{\frac{1}{2}} \tag{23}$$

The upper and lower bounds of the quadratic function can be used for a general function as long as the discrete step size (for discrete optimization) is much smaller than unity. In this case the quadratic function represents a good approximation to the general function. Also, if we use Quasi-Newton methods for optimization the inverse of the Hessian matrix is available at the optimum and this can be used in (23).

Constrained Optimization. Let us assume that we have an integer point $\bar{\underset{\sim}{x}}^0$ which is feasible. The optimum integer solution must satisfy the original constraints plus an additional constraint which is

$$g_{m+1}\ (\underset{\sim}{x}) = F(\bar{\underset{\sim}{x}}^0) - F(\underset{\sim}{x}) \geq 0 \tag{24}$$

Therefore the optimum integer solution must lie inside the region R where

$$R \triangleq \{\underset{\sim}{x} | g_i(x) \geq 0 \ , \ i = 1, 2, \ldots, m + 1\} \tag{25}$$

Again, the upper and lower bounds on the parameters can be found as in the unconstrained case.

Let us consider the following example.

$$\text{minimize} \quad (x_1 - 8)^2 + (x_2 - 2)^2 \tag{26}$$

Subject to

$$g_1 = x_2 - 0.1x_1^2 \geq 0 \tag{27}$$

$$g_2 = -x_2 - \frac{1}{3}x_1 + 4.5 \geq 0 \tag{28}$$

plus the additional constraint that x_1 and x_2 are integers. The problem is shown geometrically in Figure 2. Let us assume that $\bar{\underset{\sim}{x}}^0 = [1\ 2]^T$, which is a feasible point. Then the additional imposed constraint is

$$g_3 = 49 - (x_1 - 8)^2 - (x_2 - 2)^2 \geq 0 \tag{29}$$

From Figure 2 it can be seen that the region $\bar{R}_1$ is given by

$$\bar{R}_1 = \{\underset{\sim}{x} | 1 \leq x_1 \leq 5 \ , \ 1 \leq x_2 \leq 4\}$$

and the integer optimum solution is $\bar{\underset{\sim}{x}} = [4\ 2]^T$.

CONCLUSIONS

Some theoretical ideas for discrete nonlinear programming for which could be found practical implementation were discussed. It was shown how to find upper and lower bounds on the parameters, which form a hypercube inside of which the discrete optimum will lie. The next stage of research consists of the following two steps: (a) Based on the information of the gradient vector, to find lower and upper bounds on the parameters without performing 2n optimizations, (b) Once the hypercube is constructed, to find an efficient method to locate the discrete optimum without calculating the value of the objective function at all points inside the hypercube.

ACKNOWLEDGEMENT

The author wishes to thank Dr. J.W. Bandler of the Department of Electrical Engineering, McMaster University, Hamilton, Ontario, Canada, Dr. R. Burns and M. Carter of the Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada for useful discussions. This work was supported by the National Research Council of Canada, through a Postdoctorate Fellowship to the author.

REFERENCES

[1] T.C.Hu, "Integer Programming and Network Flows", Addison-Wesley, 1969.

[2] R.S.Garfinkel and G.L. Nemhouser, "Integer Programming", John Wiley & Sons, 1972.

[3] K. Steiglitz, "Designing Short-Word Recursive Digital Filters", Proc. 9th Allerton Conference Circuit and Systems Theory (Urbana Ill), October 1971, pp. 778-788.

[4] M. Suk and S.K. Mitra, "Computer-Aided Design of Digital Filters with Finite Word Lengths", IEEE Trans. Audio Electroaccoustics, Vol. AU-20 December 1972, pp. 356-363.

[5] J.W. Bandler and C. Charalambous, "Practical least pth Optimization of Networks", IEEE Trans. Microwave Theory Tech., vol. MTT-20, December 1972, pp. 834-840.

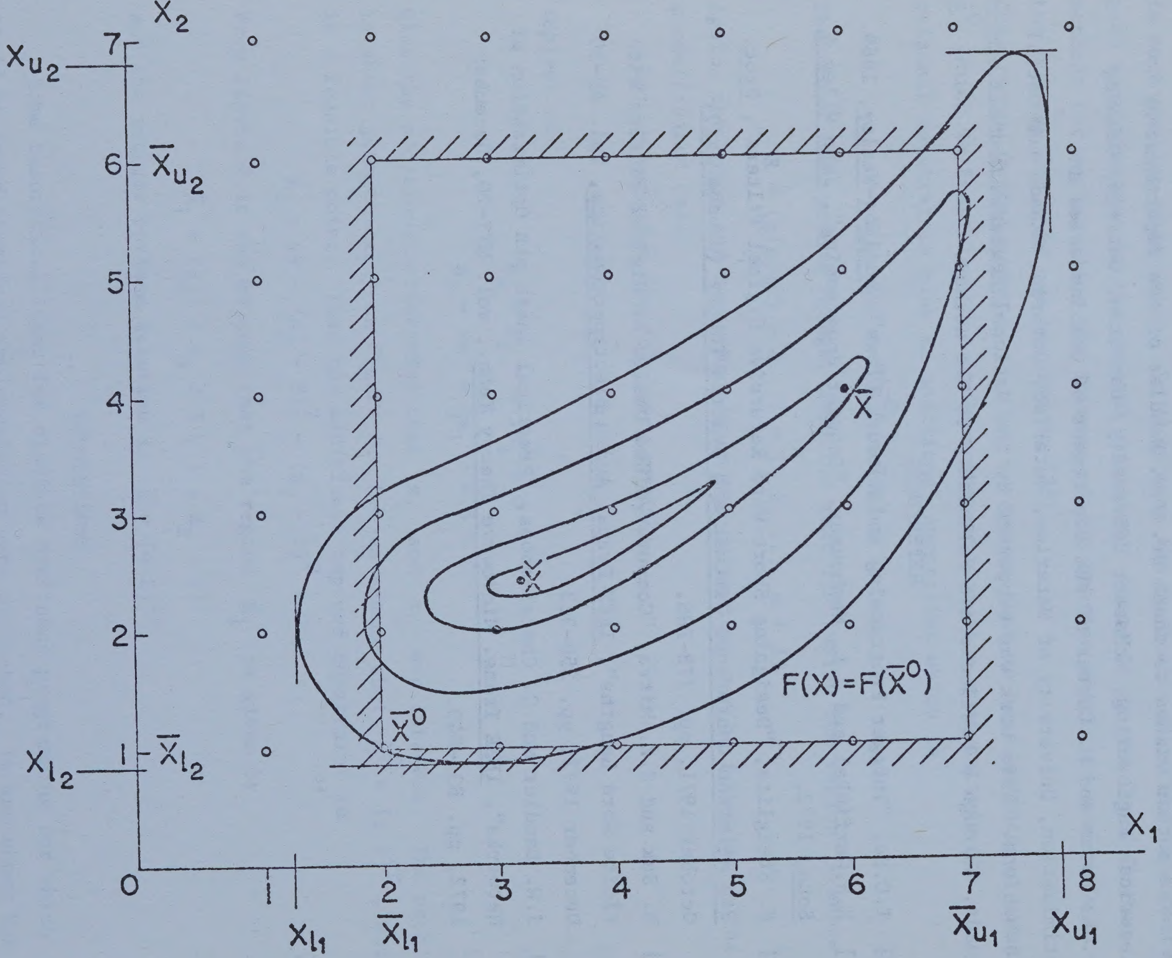

X_2
X_{u_2}
$\overline{X}_{u_2}$
X_{l_2}
$\overline{X}_{l_2}$
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
X_1
X_{l_1}
$\overline{X}_{l_1}$
$\overline{X}_{u_1}$
X_{u_1}
$\overline{X}$
$\check{X}$
$\overline{X}^0$
$F(X)=F(\overline{X}^0)$

Figure 1.
Illustration of region R_1 .

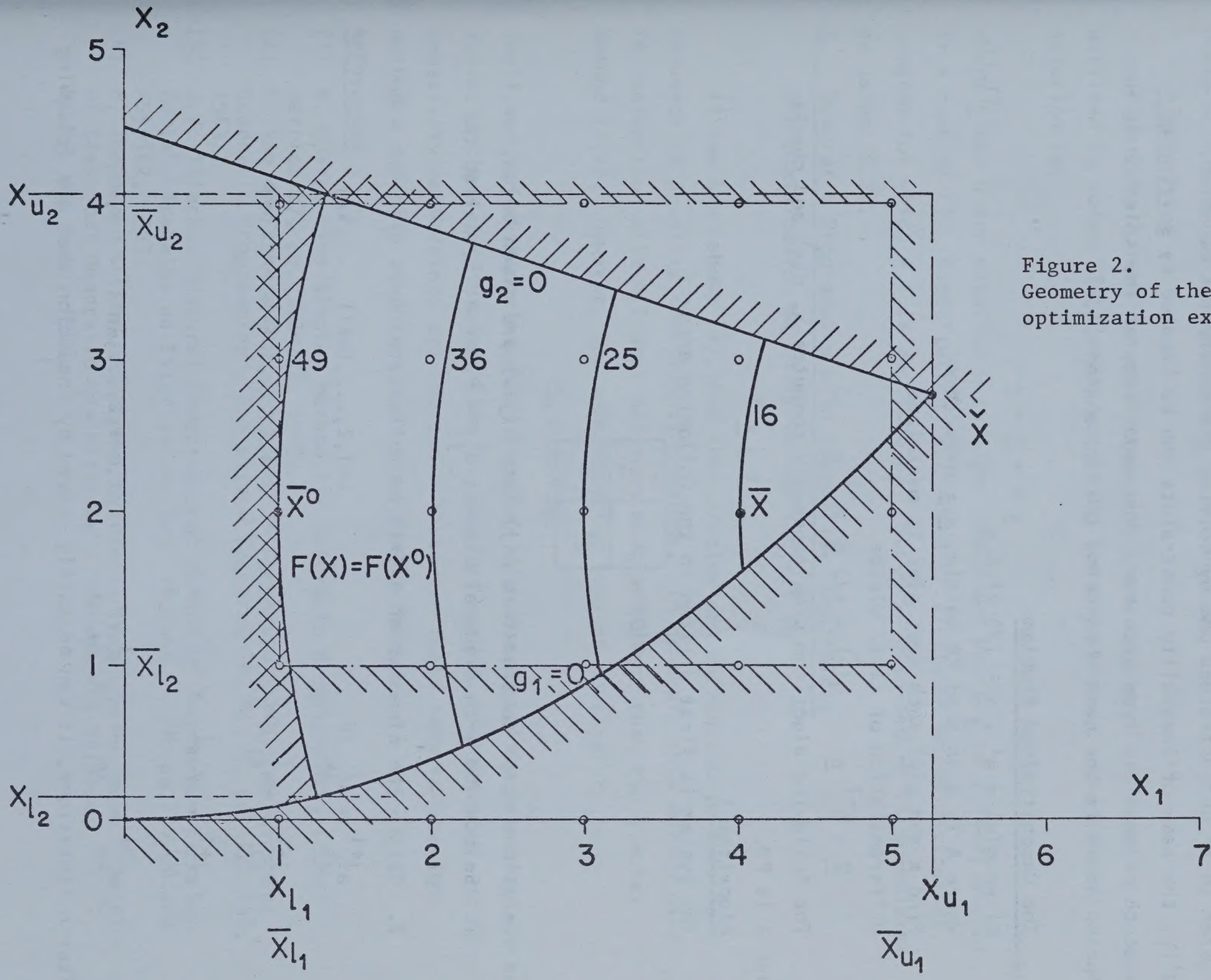

X2
5
4
3
2
1
0
Xu2
X̄u2
X̄l2
Xl2
g2=0
g1=0
49
36
25
16
X̄0
X̄
X̌
F(X)=F(X0)
X1
1
2
3
4
5
6
7
Xl1
X̄l1
Xu1
X̄u1

Figure 2.
Geometry of the constrained optimization example.

APPLICATION OF THE SYMMETRIC ALGORITHM TO SOLVE QUADRATIC PROGRAMMING PROBLEMS

J.P. COMBOT AND S.H. WU
Department of Electrical Engineering
Marquette University, Milwaukee, Wisconsin

Abstract

Two algorithms to solve the QP problems are presented. The first algorithm deals with the unconstrained problem and the second reduces the constraint problem to an unconstrained one by solving the constraint equation. Finally, the case of inequality constraints can be treated by solving a sequence of equality type problems. Non-negativity of the solution is not imposed as it is the case when using the Lagrangian [1,2].

1. The Unconstrained Problem

Given $g(\underline{x}) = \underline{a}' \cdot \underline{x} - 1/2\ \underline{x}'\ A\ \underline{x}$ (1.1)

Where A is a (n x n) PD matrix, $\underline{a},\underline{x}$ n-vectors

Find $\underline{x}$ and $g(\underline{x})$ such that $g(\underline{x})$ is maximized.

Differentiation of (1.1) yields

$$\hat{\underline{x}} = A^{-1}\ \underline{a} \qquad g(\hat{\underline{x}}) = 1/2\ \underline{a}'\ A^{-1}\ \underline{a}$$

The following algorithm simultaneously computes $\hat{\underline{x}}$, $g(\hat{\underline{x}})$ and checks that A is PD.

Algorithm 1

The system is first written in the following array

$$A(0) = \begin{array}{|c|c|} \hline A & \underline{a} \\ \hline \underline{a}' & o \\ \hline \end{array}$$

The iterations steps to transform A(j) into A(j+1) are given below,

1. Select the j-th diagonal element $a^j_{jj} = P_j$ of A(j), called the pivot.
2. Obtain the elements of A(j+1) as follows,

$$a^{j+1}_{jk} = a^j_{jk}\ /P_j \qquad k=1,2,\ldots,(n+1) \qquad k\neq j$$

$$a^{j+1}_{\ell j} = -a^j_{\ell j}\ /P_j \qquad \ell=1,2,\ldots,(n+1) \qquad p\neq j$$

$$a^{j+1}_{jj} = 1/\ a^j_{jj}$$

$$a^{j+1}_{\ell m} = a^j_{\ell m} - a^j_{jm}\ a^j_{\ell j}/P_j \qquad \ell,m=1,2,\ldots,(n+1) \qquad \ell,m\neq j$$

After n iterations, it can be easily proved by induction that the following

$A(n)$ is obtained,

$$A(n) = \left[\begin{array}{c|c} A^{-1} & A^{-1}\underline{a} \\ \hline -\underline{a}'A^{-1} & -\underline{a}'A^{-1}\underline{a} \end{array}\right]$$

Furthermore, A is PD if and only if $P_j > 0$ for $j=1,2,\dots,n$

2. Quadratic Program With Equality Constraints

The maximization of (1.1) is now subject to

$$C\,\underline{x} = \underline{f} \tag{2.1}$$

where C is a (k x n) matrix and $\underline{f}$ is a k-vector.

Algorithm 2

Algorithm 2 uses the same iteration steps as Algorithm 1 and is utilized to solve (2.1) while verifying its consistency. The general solution is,

$$\underline{x} = \underline{F} + G\,\underline{y} \tag{2.2}$$

where $\underline{F}$ and G are obtained with Algorithm 2 and where $\underline{y}$ is a (n-r) vector (r = rank of C). Substituting (2.2) into (1.1) yields an unconstrained program for which A.1 is applicable. Once $\hat{\underline{y}}$ is obtained, $\hat{\underline{x}}$ is calculated by using (2.2).

3. Quadratic Programming with Inequality Constraints

The maximization of (1.1) is now subject to

$$C\,\underline{x} \leq \underline{f}$$

It has been shown [4], that this problem can be reduced to solving a sequence of equality constraints problems. In general, this method used in conjunction with A.2 does not seem more efficient than the simplex method [1,2]. However, for some specific type of constraints, e.g.

$$\underline{x}_m \leq \underline{x} \leq \underline{x}_M$$

Theil and Van de Panne's method together with the algorithm presented requires 9 times less storage than the simplex method. Furthermore, no non-negativity constraints are imposed. These two facts seem to give the method a definite advantage over the simplex method.

References

[1] P. Wolfe, "The Simplex Method for Quadratic Programming", Econometrica 27, pp. 382-398, 1959.

[2] C. Van de Panne and A. Whinston, "A Comparison of Two Methods for Quadratic Programming", Operations Research 14, No. 3, pp. 422-441, 1963.

[3] J.V. Talacko, "Natural Computational Method of Two-Person Zero-Sum Games", Trabajos de Estadistica, Vol. 14, Nos. 1-11, pp. 17-60 and 159-162, 1963.

[4] H. Theil and C. Van de Panne, "Quadratic Programming as an Extension of Classical Quadratic Maximization", Management Science, Vol. 7, No. 1, pp. 1-20, 1960.

SNAPSHOT SCHEDULING RULES FOR SERVICE SYSTEMS WITH NONLINEAR COSTS

P. B. HENDERSON AND K. STEIGLITZ
Detartment of Electrical Engineering
Princeton University
Princeton, N.J. 08540

Abstract

The implementation of non-preemptive scheduling rules for systems with a single server and nonlinear waiting time costs is considered. In particular, the class of rules is restricted to those which select the next job on the basis of a static optimization problem: that of minimizing the total cost of the jobs remaining in the system(snapshot scheduling). Necessary conditions for the optimal sequencing of jobs are given for convex non-decreasing and quadratic waiting time costs. These conditions and a new lower bound function are used in a branch-and-bound algorithm to obtain computational results.

Introduction

Service systems which meet specified constraints are often required. For example, a trade-off between the mean waiting time in the system and the variance of the waiting time in the system may be desired. Conventional scheduling rules(i.e. first-come, first-served, etc.) are generally easy to implement; however, resulting system performance may be unacceptable. A more flexible service system may be obtained by permitting the waiting time cost to be a nonlinear function of the actual waiting time and scheduling to minimize the mean cost of operating the system. Schrage [6] presents an attractive heuristic scheduling rule for waiting time costs which are quadratic. His simulation results indicate some success in "balancing" several system parameters, the most notable being the variance of the waiting time and the mean waiting time. The purpose of this paper is to investigate the implementation of a general class of scheduling rules ("snapshot scheduling rules") for single-server service systems with nonlinear waiting time costs. Several common assumptions and definitions are now presented.

Consider a non-preemptive service system which commences operation at time $t = 0$ with no jobs in the system. Job arrivals are determined by a general input process $\{Ai\}$ with mean arrival rate $\lambda < \infty$. For job Ji arriving at time $Ai \geq 0$, let $Pi > 0$ be the service time, $Di \geq Ai$ the departure time from the queue, $Wi = Di - Ai \geq 0$ the total resident time in the queue and $Fi = Wi + Pi$ the total time in the system. When a job departs from the

---

This work was supported by NSF Grant GJ-965 and the U.S. Army Research Office(Durham) under Contract DAHCO4-69-C-0012.

system at time $t > 0$ with $n \geq 1$ jobs remaining in the queue, a scheduling rule specifies the next job to be serviced. The ultimate objective is the minimization of the long-run expected average cost per job [7]. For example, with random arrivals and minimization of the average time in the system as the objective, the optimal rule is to service the job with the shortest service time [1]. Unfortunately, for an arbitrary nonlinear cost function f(.), there are no results relating our stated objective and the required optimal rule. An intuitively appealing rule is to choose the initial job of a job sequence(commencing at departure time t) which minimizes the objective function $\{\Sigma_{i=1}^{n} f(W[i])\}$ or $\{\Sigma_{i=1}^{n} f(F[i])\}$, where [i] represents the i th job in the sequence. We call such a rule a "snapshot scheduling rule", and investigate the implementation of this scheduling rule for (1) various classes of nonlinear functions f(.), (2) degenerate service time distributions, and (3) general service time distributions. One global assumption is that all cost functions are members of the class of regular measures of performance. Hence, for the snapshot sequencing problem commencing at time t, schedules involving inserted idle-time need not be considered [1]. The regularity requirement will be met when the nonlinear waiting time cost function f(.) is monotone nondecreasing [3].

Service Systems with Degenerate Service Time Distributions

Consider a service system in which all jobs entering the system have identical service times and the nonlinear waiting time cost function f(.) is a convex non-decreasing (CND) function. It will be shown that for snapshot scheduling, first-come, first-served (FCFS) sequencing will always minimize either objective function $\{\Sigma_{i=1}^{n} f(W[i])\}$ or $\{\Sigma_{i=1}^{n} f(F[i])\}$. Before proceeding, a preparatory lemma is presented.

Lemma 1. Let f(x) be a convex non-decreasing function and X1, X2, X3, and X4 be four arguments which satisfy the inequalities:

(i) $X3 \leq \min(X1,X2,X4)$
(ii) $X4 \geq \max(X1,X2,X3)$
(iii) $X1 - X3 \leq X4 - X2$, then

$$f(X3) + f(X4) \geq f(X1) + f(X2).$$

Proof. The result is obviously true if $X3 = X4$, hence assume $X3 \neq X4$.

Let: $\alpha 1 = (X1 - X3)/(X4 - X3)$
$\alpha 2 = (X4 - X2)/(X4 - X3)$, where $0 \leq \alpha 1 \leq \alpha 2 \leq 1$.

The following inequalities are valid for any convex function f(.) and inequalities (i) and (ii):

(a) $\alpha 2 f(X3) + (1 - \alpha 2) f(X4) \geq f(X2)$
(b) $(1 - \alpha 1) f(X3) + \alpha 1 f(X4) \geq f(X1)$.

Thus,

$$\{f(X3) + f(X4) + (\alpha 2 - \alpha 1)(f(X3) - f(X4))\} \geq \{f(X1) + f(X2)\}.$$

But since:

(c) $\alpha 2 - \alpha 1 \geq 0$
(d) $f(X3) - f(X4) \leq 0$

the required inequality follows. □

The first theorem, presented below, is more general than the result required for snapshot sequencing rules. Here the time t at which sequencing may commence is permitted to be $0 \leq t < oo$. Hence for a set of n jobs $J1, J2, \ldots, Jn$ with identical service times($P > 0$), and arrival times $0 \leq A1 < A2 < \ldots < An < oo$, if $t \leq A1$, then the problem of sequencing $J1, J2, \ldots, Jn$ with apriori knowledge of service times and arrival times is being considered. For job Ji, let Di be the time at which service begins; then define $Wi = Di - Ai$ and $Fi = Wi + P$. A sequence $J[1]J[2]\ldots J[n]$ is *realizable* if and only if $W[i] \geq 0$, $i = 1, 2, \ldots, n$. Now consider two jobs $J[k]$ and $J[l]$ of a realizable sequence $J[1]J[2]\ldots J[n]$. These two jobs are called *interchangable* if the new sequence generated by interchanging $J[k]$ and $J[l]$ is realizable.

Theorem 1. Let $f(.)$ be a convex non-decreasing function and $0 \leq t < oo$. If the set of jobs $J1, J2, \ldots, Jn$, having identical service time P, are sequenced FCFS commencing at time t, then the objective function $\{\Sigma_{i=1}^{n} f(W[i])\}$ will be minimized.

Proof. Let $A1, A2, \ldots, An$ be the arrival times of jobs $J1, J2, \ldots, Jn$. Assume $J[1]J[2]\ldots J[n]$ is an optimal realizable sequence starting at time t containing at least one pair of jobs $J[k]$ and $J[l]$ for which $k < l$ and $A[l] < A[k]$. Now, $A[l] < A[k] \leq D[k] < D[l]$, hence $J[k]$ and $J[l]$ are interchangable.

Let: $W[k] = D[k] - A[k] \geq 0$

$\quad W[l] = D[l] - A[l] > 0$.

Interchange $J[k]$ and $J[l]$, then let:

$\quad W[k]' = D[l] - A[k] > 0$, new waiting time of $J[k]$

$\quad W[l]' = D[k] - A[l] > 0$, new waiting time of $J[l]$.

Note that since the service times of $J[k]$ and $J[l]$ are identical, only the waiting times of jobs $J[k]$ and $J[l]$ are affected by this interchange.

Since $A[l] < A[k]$ and $k < l$, then $W[l] > W[k]$; also, $k < l$ implies $W[k]' > W[k]$ and $W[l]' < W[l]$. Now, $A[k] - A[l] = W[l]' - W[k] = W[l] - W[k]' > 0$; hence $W[l]' > W[k]$ and $W[l] > W[k]'$.

Thus, the following inequalities are obtained:

(i) $W[l] > \max(W[l]', W[k]', W[k])$

(ii) $W[k] < \min(W[l]', W[k]', W[l])$

(iii) $W[l]' - W[k] = W[l] - W[k]'$.

Therefore, by Lemma 1,

$\{f(W[l]) + f(W[k])\} \geq \{f(W[l]') + f(W[k]')\}$, which is a contradiction to the assumed optimality of the sequence $J[1]J[2]\ldots J[n]$. □

A similiar result follows for the system waiting time.

Corollary 1. Likewise, under the conditions of the theorem, FCFS sequencing minimizes $\{\Sigma_{i=1}^{n} f(F[i])\}$.

Corollary 2. Consider a scheduling discipline with k distinct service times $P1, P2, \ldots, Pk$. A necessary condition for a schedule to minimize $\{\Sigma_{i=1}^{n} f(W[i])\}$ or $\{\Sigma_{i=1}^{n} f(F[i])\}$ when $f(.)$ is CND, is that the jobs in each service time class are sequenced FCFS.

Proof. Note that when interchanging two jobs in the same service time class, only the waiting time of these two jobs are affected. $\square$

Thus, for any finite set of jobs with identical service times and specified arrival times, the total cost when f(.) is CND will be minimized by first-come, first-served sequencing with no inserted idle time.

Snapshot Scheduling with General Service Time Distribution

The implementation of snapshot scheduling rules for a service system where the waiting time function f(.) is convex non-decreasing (CND) and service times are arbitrary is considered. If, at departure epoch t, n $\geq$ 1 jobs occupy the queue, a sequence of jobs J[1]J[2]...J[n] starting at time t for which $\{\Sigma_{i=1}^{n} f(W[i])\}$ or $\{\Sigma_{i=1}^{n} f(F[i])\}$ is minimized will specify the next job for service. If jobs Jk and Jl have identical service times, then sequencing Jk and Jl FCFS is a necessary condition for optimality by Corollary 2, Theorem 1. Now similiar conditions are derived for jobs with distinct service times. Let Pi be the service time for job Ji.

Lemma 2. Let J[k] and J[l] be two jobs in a job sequence J[1]J[2]...J[n], commencing at time t, which satisfy the conditions (a) k < l, (b) P[k] $\leq$ P[l] and (c) A[k] $\leq$ A[l]. Then, interchanging J[k] and J[l] cannot decrease the objective function $\{\Sigma_{i=1}^{n} f(W[i])\}$, where f(.) is CND.

Proof. Let R[i] = t - A[i] $\geq$ 0 be the resident time of job J[i] in the system at departure epoch t.
The inequality

$$\{f(R[l]+C) + f(R[k+1]+C+P[l]) + \ldots + f(R[k]+C+\Sigma_{i=k+1}^{l} P[i])\} \geq$$

$$\{f(R[k]+C) + f(R[k+1]+C+P[k]) + \ldots + f(R[l]+C+\Sigma_{i=k}^{l-1} P[i])\},$$

where $C = \Sigma_{i=1}^{k-1} P[i]$, must be shown.

Since f(.) is non-decreasing and P[k] $\leq$ P[l],

$$f(R[k+j]+C+\Sigma_{i=k+1}^{j-1} P[i]+P[l]) \geq f(R[k+j]+C+\Sigma_{i=k}^{j-1} P[i]),$$

j = 1, 2, ..., l-k-1. Thus, only the following inequality need be considered,

$$\{f(R[l]+C) + f(R[k]+C+\Sigma_{i=k+1}^{l} P[i])\} \geq$$

$$\{f(R[k]+C) + f(R[l]+C+\Sigma_{i=k}^{l-1} P[i])\}.$$

To simplify notation let:

(a) $X1 = R[l] + C + \Sigma_{i=k}^{l-1} P[i]$

(b) $X2 = R[k] + C$

(c) $X3 = R[l] + C$

(d) $X4 = R[k] + C + \Sigma_{i=k+1}^{l} P[i]$.

Since R[l] $\leq$ R[k] and P[k] $\leq$ p[l] the inequalities (i)-(iii) below are binding:

(i) $X3 \leq \min(X1, X2, X4)$
(ii) $X4 \geq \max(X1, X2, X4)$.

Let: $\alpha 1 = \Sigma_{i=k}^{l-1} P[i]/(X4 - X3) = (X1 - X3)/(X4 - X3)$.

$\alpha 2 = \Sigma_{i=k+1}^{l} P[i]/(X4 - X3) = (X4 - X2)/(X4 - X3)$.
Note: $X3 \neq X4$ since $P[l] > 0$.

(iii) $\alpha 1 \leq \alpha 2$.

Thus by Lemma 1, $\{f(X3) + f(X4)\} \geq \{f(X1) + f(X2)\}$. □

Lemma 3. Let J[k] and J[l] be two jobs in a sequence J[1]J[2]...J[n], commencing at time t, which satisfy the conditions (a) $k < l$, (b) $P[k] \leq P[l]$, (c) $A[l] \leq A[k]$ and (d) $A[k] + P[k] \leq A[l] + P[l]$. Then, interchanging J[k] and J[l] cannot decrease the objective function $\{\Sigma_{i=1}^{n} f(W[i])\}$, where f(.) is non-decreasing.

Proof. From Lemma 2, only the following inequality need be shown:

$\{f(R[l] + C) + f(R[k] + C' + P[l])\} \geq$
$\{f(R[k] + C) + f(R[l] + C' + P[k])\}$, where

$C' = C + \Sigma_{i=k+1}^{l-1} P[i]$ and hence $C' \geq C$.

Now, with $R[l] \geq R[k]$,
(i) $f(R[l] + C) \geq f(R[k] + C)$.
(ii) $f(R[k] + C' + P[l]) \geq f(R[l] + C' + P[k])$ if $R[k] + P[l] \geq R[l] + P[k]$, since f(.) is non-decreasing. □

Now Lemma 2 and Lemma 3 imply:

Theorem 2. Let J[k] and J[l] be two jobs in a sequence J[1]J[2]...J[n], which satisfy the conditions (a) $k < l$, (b) $P[k] \leq P[l]$ and (c) $A[k] + P[k] \leq A[l] + P[l]$. Then, interchanging J[k] and J[l] cannot decrease $\{\Sigma_{i=1}^{n} f(W[i])\}$, where f(.) is CND.

A similiar result follows for the system waiting time.

Corollary 1. Replace condition (c) with $A[k] \leq A[l]$; then, interchanging J[k] and J[l] cannot decrease $\{\Sigma_{i=1}^{n} f(F[i])\}$, where f(.) is CND.

Only conditions under which interchanging two jobs will not decrease the value of the objective function are given. If the conditions of the theorem are not met, then the value of the objective function may either increase or decrease upon interchanging two jobs. Unfortunately, these results are not usually sufficient to construct an optimal sequence, and some stronger sequencing technique is required to generate exact solutions.

Application of the Branch-and-Bound Algorithm to Optimal Sequencing

One sequencing technique which yields exact solutions is the branch-and-bound algorithm. In general this algorithm is computationally expensive but often the only technique available for obtaining exact solutions. In addition, exact solutions are useful for measuring the effectiveness of heuristic sequencing methods. In this section a useful lower bound function will be derived for finding an optimal snapshot sequence, where $f(.)$ is CND.

Assume $m < n$ jobs have been scheduled (i.e. a partial permutation of m jobs) at a designated node πx of the branch-and-bound algorithm. Let $COST(\pi x) = \Sigma_{i=1}^{m} f(W[i])$, where $J[1]J[2]...J[m]$ is the sequence of m jobs specified by πx. Consider the following function $L(\pi x)$ associated with πx.

$$L(\pi x) = COST(\pi x) + (n - m) f(\Sigma_{i=m+1}^{n} W[i]spt/(n - m)), \text{ where}$$

$W[i]spt$ indicates the remaining $(n - m)$ jobs are sequenced shortest-processing-time-first.

A useful lower bound function satisfies the following properties [3]:

(i) If πx is any node and πy a complete solution, then $L(\pi x) \leq L(\pi y)$.

(ii) If πy represents a complete solution, then $L(\pi y) = COST(\pi y)$.

Trivally, (ii) holds when $n = m$. Note also that if πx represents a partial permutation of $(n - 1)$ jobs, then $L(\pi x) = COST(\pi y)$, where πy is the complete solution. Theorem 3 confirms property (i).

Theorem 3. Let πx be a partial permutation of $m < n$ jobs; then $L(\pi x) \leq \Sigma_{i=1}^{n} f(W[i]')$, where $f(.)$ is CND and,

(a) $[i]' = [i]$, $i = 1,2,...,m$
(b) $[i]'$ is arbitrary, $i = m+1,...,n$.

Proof. The proof of the theorem proceeds by showing:

(i) $\min\{1/(n-m) \Sigma_{i=m+1}^{n} W[i]\} = 1/(n-m) \Sigma_{i=m+1}^{n} W[i]spt$, and

(ii) $f(1/(n-m) \Sigma_{i=m+1}^{n} W[i]spt) \leq \min\{1/(n-m) \Sigma_{i=m+1}^{n} f(W[i])\}$, where the minimization in each case is over all sequences of $(n-m)$ remaining jobs.

Proof of (i): Let Di be the time at which Ji commences service. Now,

$$\Sigma_{i=m+1}^{n} W[i] = \Sigma_{i=m+1}^{n} (D[i] - A[i]) = \Sigma_{i=m+1}^{n} D[i] - \Sigma_{i=m+1}^{n} A[i].$$

Since $\Sigma_{i=m+1}^{n} A[i]$ is constant for all job sequences, the result follows from the optimality of SPT sequencing with concurrent arrivals [1].

Proof of (ii): Let $J[m+1]'J[m+2]'...J[n]'$ be a sequence which minimizes the right side of (ii).

From (i)

$$\Sigma_{i=m+1}^{n} W[i]spt \leq \Sigma_{i=m+1}^{n} W[i]'.$$

Since $f(.)$ is nondecreasing,

$$f(1/(n-m) \Sigma_{i=m+1}^{n} W[i]spt) \leq f(1/(n-m) \Sigma_{i=m+1}^{n} W[i]').$$

But, $f(1/(n-m) \Sigma_{i=m+1}^{n} W[i]') \leq (1/(n-m) \Sigma_{i=m+1}^{n} f(W[i]')$ follows directly from the convexity condition for convex function $f(.)$. □

Corollary 1. $\Sigma_{i=1}^{m} f(F[i]) + (n-m) f(1/(n-m) \Sigma_{i=m+1}^{n} F[i]spt)$ is a lower bound function for the objective function $\{\Sigma_{i=1}^{n} f(F[i])\}$.

Thus for the general CND snapshot sequencing problem, Theorem 3 gives a lower bound function and Theorem 2 may be used to eliminate non-optimal sequences from consideration for a branch-and-bound algorithm. When the class of waiting time functions f(.) is further restricted, it is generally possible to obtain stronger results as is illustrated now for a class of quadratic functions.

Restrict f(.) to be CND and Quadratic(non-neg. arguments)

Many specific and useful results may be obtained by the further restriction of f(.) to be a convex non-decreasing quadratic function for non-negative arguments. Perhaps one of the most important of these is the relationship between optimal snapshot sequences for minimizing the waiting time cost of the queue and the waiting time cost of the system.

Lemma 4. For the snapshot sequencing problem, J[1]J[2]...J[n] is a sequence for which $\{\Sigma_{i=1}^{n} (\alpha W[i] + \beta W[i]^2)\}$ is minimized if and only if it is a sequence for which $\{\Sigma_{i=1}^{n} (\alpha F[i] + \beta F[i]^2)\}$ is minimized.

Proof. It will be shown that for every sequence of jobs these two objective values differ by a fixed constant. To simplify the proof the following equalities will be shown independently. Let J[1]J[2]...J[n] and J[1]'J[2]'...J[n]' be two arbitrary sequences. Then,

(i) $\Sigma_{i=1}^{n} W[i] - \Sigma_{i=1}^{n} W[i]' = \Sigma_{i=1}^{n} F[i] - \Sigma_{i=1}^{n} F[i]'$

(ii) $\Sigma_{i=1}^{n} W[i]^2 - \Sigma_{i=1}^{n} W[i]'^2 = \Sigma_{i=1}^{n} F[i]^2 - \Sigma_{i=1}^{n} F[i]'^2$.

Proof of (i). $\Sigma_{i=1}^{n} W[i] - \Sigma_{i=1}^{n} W[i]' =$

$\Sigma_{i=1}^{n} (R[i] + \Sigma_{j=1}^{i-1} P[j]) - \Sigma_{i=1}^{n} (R[i]' + \Sigma_{j=1}^{i-1} P[j]') =$

$\Sigma_{i=1}^{n} \Sigma_{j=1}^{i-1} (P[j] - P[j]') =$

$\Sigma_{i=1}^{n} \Sigma_{j=1}^{i-1} (P[j] - P[j]') + \Sigma_{i=1}^{n} (P[i] - P[i]') =$

$\Sigma_{i=1}^{n} \Sigma_{j=1}^{i} (P[j[- P[j]') =$

$\Sigma_{i=1}^{n} (R[i] + \Sigma_{j=1}^{i} P[j]) - \Sigma_{i=1}^{n} (R[i]' + \Sigma_{j=1}^{i} P[j]') =$

$\Sigma_{i=1}^{r} F[i] - \Sigma_{i=1}^{n} F[i]'$.

Proof of (ii). $\Sigma_{i=1}^{n} F[i]^2 - \Sigma_{i=1}^{n} F[i]'^2 =$

$\Sigma_{i=1}^{n} (R[i] + \Sigma_{j=1}^{i} P[j])^2 - \Sigma_{i=1}^{n} (R[i]' + \Sigma_{j=1}^{i} P[j]')^2 =$

$\Sigma_{i=1}^{n} \{(R[i] + \Sigma_{j=1}^{i-1} P[j]) + P[i]\}^2 - \Sigma_{i=1}^{n} \{(R[i]' + \Sigma_{j=1}^{i-1} P[j]') + P[i]'\}^2 =$

$$\{\Sigma_{i=1}^{n}(R[i]+\Sigma_{j=1}^{i-1}P[j])^2+2[\Sigma_{i=1}^{n}P[i]R[i]+\Sigma_{i=1}^{n}P[i]\Sigma_{j=1}^{i-1}P[j]]+\Sigma_{i=1}^{n}P[i]^2\} -$$
$$\{\Sigma_{i=1}^{n}(R[i]'+\Sigma_{j=1}^{i-1}P[j]')^2+2[\Sigma_{i=1}^{n}P[i]'R[i]'+\Sigma_{i=1}^{n}P[i]'\Sigma_{j=1}^{i-1}P[j]'] +$$
$$\Sigma_{i=1}^{n}P[i]'^2\}.$$

Clearly:

(a) $\Sigma_{i=1}^{n} P[i]R[i] = \Sigma_{i=1}^{n} P[i]'R[i]'$

(b) $\Sigma_{i=1}^{n} P[i]^2 = \Sigma_{i=1}^{n} P[i]'^2$.

Also,

(c) $\Sigma_{i=1}^{n} P[i]\Sigma_{j=1}^{i-1} P[j] = \Sigma_{i=1}^{n} P[i]'\Sigma_{j=1}^{i-1} P[j]'$.

This is seen by noticing that every possible product occurs exactly once.

Thus finally this yields,

$$\Sigma_{i=1}^{n} F[i]^2 - \Sigma_{i=1}^{n} F[i]'^2 =$$
$$\Sigma_{i=1}^{n} (R[i] + \Sigma_{j=1}^{i-1} P[j])^2 - \Sigma_{i=1}^{n} (R[i]' + \Sigma_{j=}^{i-} P[j]')^2 =$$
$$\Sigma_{i=1}^{n} W[i]^2 - \Sigma_{j=1}^{n} W[i]'^2. \ \square$$

Therefore, it is only necessary to consider the minimization with respect to $\{\Sigma_{i=1}^{n} (\alpha W[i] + \beta W[i]^2)\}$ since, in general, during analysis this function yields less complicated expressions. Unfortunately, this theorem cannot be extended to include the class of CND functions f(.). It is a trival exercise to construct a counterexample for a two job sequence.

A second interesting aspect of the quadratic function is the variety of the results which may be derived. However, to initiate this investigation, a counterexample to a representative conjectured relationship will be presented. It is perhaps noteworthy that extensive computational tests were unable to generate a counterexample, indicating the usefulness of this conjecture as a heuristic.

Conjecture 1. Let f(.) be quadratic nondecreasing for non-negative arguments. Consider a snapshot scheduling problem commencing at time t to minimize $\{\Sigma_{i=1}^{n} (\alpha W[i] + \beta W[i]^2)\}$. If for two jobs Jk and Jl, $Pk \le Pl$ and COST(JKJL) < COST(JLJK), then a necessary condition for optimality is that Jk precede Jl.

Counterexample 1. Let $\alpha = 0$, $\beta = 1$, $n = 3$ and for the three jobs Jk, Jh, Jl let:

(i) $Pk = 20$, $Rk = 126$

(ii) $Ph = 9$, $Rh = 0$

(iii) $Pl = 112$, $Rl = 1000$.

Now, COST(JKJL) < COST(JLJK) and $Pk < Pl$ but the optimal sequence is JlJhJk. $\square$

The following results are stated without proof and represent several of the results found to exist for this class of quadratic functions. All results are for snapshot sequencing problems with sequencing commencing at time t.

Lemma 5. Let J[1]J[2]...J[n] be an optimal sequence. If there exist two jobs J[k] and J[l] such that (a) $l < k$, (b) $P[l] < P[k]$, then $\Sigma_{i=l-1}^{k+1} P[i] \le (k-l-1) P[k]$.

Lemma 6. If for two jobs Jk and Jl, (a) $Pk \leq Pl$ and (b) $\{2RlPk + Pk^2\} < \{2RkPk + Pl^2\}$, then a necessary condition for optimality is that Jk precede Jl.

Lemma 7. Let J[k] and J[l] be two jobs in a job sequence J[1]J[2]...J[n], commencing at time t, which satisfy the conditions (a) $k < l$,

(b) $$\{2(P[l]-P[k])\{\Sigma_{i=1}^{n}(R[i]+\Sigma_{j=1}^{i}P[j])+(l-k)C\} + (l-k-1)(P[l]^2-P[k]^2)-(2R[l]P[k]+P[k]^2)+(2R[k]P[l]+P[l]^2)\} > \{\Sigma_{i=1}^{n}P[i](R[l]-R[k])\}.$$

Then, interchanging J[k] and J[l] cannot decrease the objective function $\{\Sigma_{i=1}^{n} W[i]^2\}$, where $C = \Sigma_{i=1}^{n} P[i]$.

Preliminary Computational Results

Branch-and-bound algorithms using the lower bound function previously derived and employing several of the node elimination rules were implemented in FORTRAN IV for the objective function $\min\{\Sigma_{i=1}^{n} W[i]^2\}$. The job arrival times Ai, $i = 1,2,\ldots,n$ are independent and distributed uniformly on the interval [0,1000] and the service times Pi are independent and distributed uniformly on [0,100]. Table 1 illustrates the total number of nodes created for thirty independent snapshot sequencing problems each consisting of $n = 5,6,\ldots,10$ jobs. The elimination rules corresponding to Theorem 2, Conjecture 1 and Theorem 2 plus Lemma 7 are presented. All sequences resulting from the elimination rule based on Conjecture 1 were optimal.

	n = number of jobs sequenced					
	5	6	7	8	9	10
Theorem 2	246	740	2040	5840	13803	31056
Conj. 1	230	358	655	1486	2247	5361
Tm.2,Lm.7	226	344	571	815	1127	1595

Table 1. Total number of nodes created for 30 sequences.

REFERENCES

1. Conway, R.W., Maxwell, W.L., and Miller, L.W., Theory of Scheduling, Addison-Wesley, (1967).

2. Fox, B.L., Schrage, L.E., "The Value of Various Strategies in Branch-and-Bound," unpublished manuscript, University of Chicago, 1972.

3. Kohler, W.H., "Exact and Approximate Algorithms for Permutation Problems," Ph.D. Dissertation, Dept. of Electrical Engineering, Princeton University, August 1972.

4. Kohler, W.H., Steiglitz, K., "Characterization and Theoretical Comparison of Branch-and-Bound Algorithms for Permutation Problems," to appear JACM, Jannuary 1974.

5. Ross, S.M., Applied Probability Models with Optimization Applications, Holden-Day, (1970)

6. Schrage, L.E., "Optimal Scheduling Rules for Quadratic Waiting Costs," Seventh Annual Princeton Conference on Information Sciences and Systems, 1973.

7. Stidham, S. Jr., "Single-Server Queuing Systems," Operations Research, Vol. 18, No. 4, pp. 708-732.

OPTIMAL REPAIRMAN ASSIGNMENT

LEONARD SHAW
Polytechnic Institute of New York, Brooklyn, New York

The problem considered here assumes that manpower can be shifted between repair and maintenance assignments. In the simplest case we might consider one repairman and one maintenance man, with the latter assignable to help the former with repairs when the repair backlog is very big. The object is to minimize the mean-time to complete all repairs, subject to a penalty for the time the maintenance man is shifted away from his main duties.

It is assumed that failures occur at Poisson times, with a rate λ, and that all repair times are described by the same exponential probability law.

This work is equivalent to the control of the number of servers for a queue based on the waiting time of a new arrival. Previous work has controlled either priorities or the number of servers based on the number of customers in the queue rather than using the server's time backlog as the decision variable.

While an extra server was postulated to move in and out as the backlog rose and fell across a threshold, the best threshold appears to be zero or infinity. That is, if the extra server costs relatively little, always use him or, if he is expensive, never use him. The same results obtain for a different discipline which uses a threshold for a single choice to add a server who must then work until the end of the busy period.

These singular results are related to similar ones for vehicle queues and seem to be consequences of the linear type of cost functions.

This work was partially supported by Grant N00014-67-A-0438-0013 from the Office of Naval Research

INDEX OF AUTHORS

Adoul, J.-P. A. 616
Aggarawal, R. 693
Aldermeshian, H. 194
Amir-Ebrahimy, K. 178
Anders, J. R. 363

Baras, J. S. 1
Bingulac, S. P. 373
Bock, M. E. 324
Bona, B. E. 801
Bourquin, J. J. 758
Bracha-Barak, A. 782
Brown, C. M., Jr. 428
Brown, R. D. 115

Calahan, D. A. 729
Calise, A. J. 693
Ćalović, M. S. 373
Cantalapeidra, J. 815
Chaney, T. J. 96
Charalambous, C. 748, 840
Chyung, Dong Hak 836
Combot, J. P. 848
Corinthios, M. J. 651
Ćuk, N. M. 373

D'Angelo, H. 275
Datko, R. 11
Davidson, E. S. 547
Davis, J. H. 20
Davison, E. J. 343, 353, 790
Debs, A. S. 64
Deyst, J. 204
DuCasse, E. G. 400, 787
Duchan, A. I. 171
Dunn, K.-P. 304

Eastwood, L. F., Jr. 601
Egbert, R. I. 758
Elgerd, O. I. 45
Elhadef, D. 763
Ephremides, A. 505

Fairman, F. W. 487
Fischer, D. R. 254
Fishman, P. M. 597
Fouad, A. A. 54
Francis, B. 468
Fuhrmann, P. A. 30
Fuller, S. H. 577
Furuya, N. 106

Gault, J. W. 393
Geller, J. B. 786
Goffin, J. L. 544
Gupta, R. D. 487

Hadjiconstandis, A. 826
Hajj, I. 635
Haurie, A. 333
Helton, J. W. 131
Henderson, P. B. 850
Hexner, G. 236
Ho, C. 798
Hoppensteadt, F. 703
Hosoe, S. 477
Hostetter, G. H. 226
Huang, Hsien-Lu 495
Hughes, R. O. 453

Ito, M. 477

Jameson, A. 686
Jenkins, L. 284

Kajitani, Y. 627, 668
Kalata, P. 437
Kaliski, M. E. 786
Kashyap, R. L. 515
Koch, G. 151

Lam, N. V. 815
Larson, A. G. 547
Laurie, D. H. 568
Leake, R. J. 82
Lee, A. Y. 275
Linton, T. D. 254
Lizewski, E. J. 816
Lodi, E. 266
Luccio, F. 266
Lukes, D. L. 315

Masson, G. M. 771
McDaniel, W. L., Jr. 760
Meditch, J. S. 226
Meier, L., III 526
Metze, G. A. 400, 787
Mickle, M. H. 194
Mintz, M. 215, 236
Mitchell, J. R. 760
Monopoli, R. V. 428, 816
Moore, B., III 36
Mukherjee, S. 838
Myers, B. R. 257

Nakhla, Mrs. M. 661
Newcomb, R. W. 115

O'Malley, R. E., Jr. 678
O'Neill, W. D. 160
Ortega, J. A. 82

Parker, R. B. 187
Prasad, T. 454
Priemer, R. 437

Rabbat, N. B. 739
Rauch, R. E. 718

Reddy, P. B. 708
Reddy, S. M. 412
Rhodes, I. B. 304
Rutman, R. 247

Saeks, R. 284
Salmon, D. M. 526
Sandberg, I. W. 77
Sanders, C. W. 254
Sannuti, P. 709
Schatz, J. G. 83
Schertz, D. R. 408
Schlueter, R. A. 383
Schmitendorf, W. E. 536
Schwartz, S. C. 605
Sebakhy, O. A. 468
Seidman, T. I. 123
Sen, A. 456
Shaw, L. 860
Shida, S. 106
Shinoda, S. 668
Sims, C. S. 447
Singhal, K. 635, 661
Sinha, N. K. 454, 456, 838
Slemrod, M. 141
Smith, E. J. 763
Snyder, D. L. 597
Speyer, J. 204
Srinivasan, R. 96
Steiglitz, K. 850
Stelte, D. 215
Stephenson, C. M. 558
Stevenson, D. K. 577
Stucki, M. J. 418

Tacker, E. C. 254
Tapia, M. A. 643
Thaler, G. J. 815, 826
Thomas, J. B. 587
Traganitis, A. 587

Uno, K. 171

Vlach, J. 635, 661
Vogt, W. G. 178

Willsky, A. S. 459
Winslow, L. 284
Wonham, W. M. 468
Wood, J. R. 86, 294
Wu, S. H. 363, 848

LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.

UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
11 1973
3 0112 006279571